Novel Wavelet-Based Statistical Methods with Applications in Classification, Shrinkage, and Nano-Scale Image Analysis

A Thesis
Presented to
The Academic Faculty

by

Ilya Lavrik

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2006

Novel Wavelet-Based Statistical Methods with Applications in Classification, Shrinkage, and Nano-Scale Image Analysis

Approved by:

Dr. Brani Vidakovic, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Christopher Heil
School of Mathematics
Georgia Institute of Technology

Dr. Anthony Hayter
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yang Wang
School of Mathematics
Georgia Institute of Technology

Dr. Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: November 22, 2005

To my parents, Dasha and Heidi.

PREFACE

The subject of wavelets appeared in the mid 1980s influenced by ideas from both pure mathematics (harmonic analysis, functional analysis, approximation theory, fractal sets etc.) and applied mathematics (signal processing, mathematical physics etc.). Almost instantaneously it became a success story with thousands of papers written by now and a wide range of applications.

A simple definition of wavelets is

Definition 1 *A wavelet is a function* $\psi(t) \in \mathbb{L}_2(\mathbb{R})$ *such that the family of functions*

$$\psi_{j,k} = 2^{j/2}\psi(2^j t - k)$$

where j *and* k *are arbitrary integers, is an orthonormal basis in the Hilbert space* $\mathbb{L}_2(\mathbb{R})$.

Wavelets are widely used in many areas. This is due to their powerful properties, such as orthogonality, localization in time and frequency, compact support, as well as availability of the fast computational algorithms. Many scientists and researchers find the use of wavelets advantageous to their purposes. Naturally, this statement can be justified only after working with wavelets on some concrete problems. In this thesis, problems of classification, shrinkage/thresholding, and nano-scale image analysis are addressed, with wavelets playing a crucial role in each of them.

This thesis is organized as follows.

Chapter I contains sections listing prerequisites. We state without proof well-known results on Hilbert spaces, and Fourier transform, and give a more detailed introduction to wavelets. A reader familiar with these concepts can skip this chapter and proceed directly to Chapters II-IV.

Chapter II introduces the wavelet-based generalized linear classifier. We show that under mild conditions this classifier is consistent. The performance of the classifier is illustrated on simulated data as well as on the "real" data from the paper making process and biological example.

In Chapter III, we propose two new approaches to wavelet shrinkage/thresholding based on testing multiple hypotheses in the wavelet domain. New methods are compared to well-known thresholding

techniques and applied to atomic force microscopy data.

In Chapter IV, a new method for the analaysis of the nano-scale images is proposed. The method uses the Hough transform combined with wavelets to detect and analyze the linear structure formed by an atomic lattice. The chapter briefly introduces the Hough transform, and after description of the proposed method, shows its applications to simulated data as well as to nano-scale images of the ZnS structures.

Appendix A lists all the definitions, lemmas, and theorems which were used in the proof of the consistency in Chapter II.

Appendix B containts MATLAB codes of the Daubechies-Lagarias algorithm used in construction of the wavelet-based generalized linear classifier.

Appendix C containts the manual for the MATLAB toolbox (NSIA) in which the method proposed in Chapter IV is implemented.

ACKNOWLEDGEMENTS

I want to thank Dr. Brani Vidakovic for his assistance and encouragement.

I gratefully acknowledge Dr. Fabrizio Ruggeri and Yoon Young Jung for their contributions.

I want to thank Daniel Moore for his help and for providing images.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Given the recent popularity and clear evidence of wide applicability of wavelets, this thesis is devoted to several novel statistical applications of Wavelet transforms. Statistical multiscale modeling has, in the most recent decade, become a well established area in both theoretical and applied statistics, with impact on developments in statistical methodology.

Wavelet-based methods are important in statistics in areas such as regression, density and function estimation, factor analysis, modeling and forecasting in time series analysis, assessing self-similarity and fractality in data, and spatial statistics. In this thesis we show applicability of the wavelets by considering three problems.

- We consider a binary wavelet-based linear classifier. Both consistency results and implemental issues are addressed. We show that under mild assumptions wavelet-based classification rule is both weakly and strongly universally consistent. The proposed method is illustrated on synthetic data sets in which the "truth" is known and on applied classification problems from the industrial and bioengineering fields.

- We develop wavelet shrinkage methodology based on testing multiple hypotheses in the wavelet domain. The shrinkage/thresholding approach by implicit or explicit simultaneous testing of many hypotheses had been considered by many researchers and goes back to the early 1990's. We propose two new approaches to wavelet shrinkage/thresholding based on the local False Discovery Rate (FDR), Bayes factors and ordering of posterior probabilities.

- We propose a novel method for the analysis of straight-line alignment of features in the images based on Hough and Wavelet transforms. The new method is designed to work specifically with Transmission Electron Microscope (TEM) images taken at nanoscale to detect linear structure formed by the atomic lattice.

CHAPTER I

INTRODUCTION TO WAVELETS

1.1 Hilbert Spaces

In this chapter we state without proof well-known results on Hilbert spaces, and Fourier transform, and give a more detailed introduction to wavelets. A reader familiar with these concepts can skip this chapter and proceed directly to Chapters II-IV.

Let $\mathcal{H}$ be a linear space over either the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. An inner product on $\mathcal{H}$ is a function $\langle ., . \rangle$ from $\mathcal{H} \times \mathcal{H}$ into scalars such that

$$\langle x, y \rangle = \overline{\langle y, x \rangle}$$

$$\langle \alpha x_1 + \beta x_2, y \rangle = \alpha \langle x_1, y \rangle + \beta \langle x_2, y \rangle$$

$$\langle x, x \rangle \geq 0$$

$$\langle x, x \rangle = 0 \quad \text{if and only if} \quad x = 0.$$

In such a situation the function $\|x\| = \sqrt{\langle x, x \rangle}$ is a norm i.e. it satisfies

$$\|x + y\| \leq \|x\| + \|y\|$$

$$\|\alpha x\| = |\alpha| \|x\|$$

$$\|x\| = 0 \quad \text{if and only if} \quad x = 0.$$

A linear space $\mathcal{H}$ equipped with an inner product is a *Hilbert space* if $\mathcal{H}$ is complete as a metric space with the metric $d(x, y) = \|x - y\|$.

There are two basic examples of Hilbert spaces. For any subset $A \subset \mathbb{R}^d$, $d = 1, 2, \ldots,$ in particular for the whole $\mathbb{R}^d$ or an interval in $\mathbb{R}$, $\mathbb{L}_2(A)$ is the space of all (equivalence classes of equal a.e.) measurable functions such that

$$\|f\|_2 := \left(\int_A |f(x)|^2 \right)^{1/2} < \infty.$$

The inner product is given by

$$\langle f, g \rangle := \int_A f(x) \overline{g(x)} dx.$$

If B is a countable set then the space $\ell_2(B)$ is the space of all sequences $(a_b)_{b\in B}$ indexed by the set B such that

$$\|b\|_2 := \left(\sum_{b\in B} |a_b|^2\right)^{1/2} < \infty.$$

The inner product is given by

$$\langle a, c\rangle := \sum_{b\in B} a_b\overline{c_b}.$$

Two vectors x and y in a Hilbert space $\mathcal{H}$ are called orthogonal if $\langle x, y\rangle = 0$. Two subsets A and B of a Hilbert space $\mathcal{H}$ are orthogonal if $\langle a, b\rangle = 0$ for all $a \in A$ and $b \in B$. We denote this by $A\perp B$. A system of non-zero vectors $(x_s)_{s\in S}$ is called an orthogonal system if $\langle x_s, x_{s'}\rangle = 0$ for $s \neq s'$. If we have

$$\langle x_s, x_{s'}\rangle = \begin{cases} 0 & \text{if} \quad s \neq s' \\ 1 & \text{if} \quad s = s' \end{cases}$$

then the system is orthonormal. An orthonormal system $(x_s)_{s\in S}$ is an orthonormal basis in $\mathcal{H}$ if one of the following equivalent conditions holds:

- every $x \in \mathcal{H}$ can be written as a convergent series $x = \sum_{s\in S} a_s x_s$ for some scalars a_s
- if $\langle x, x_s\rangle = 0$ for all $s \in S$ then $x = 0$
- for every $x \in \mathcal{H}$ the series $\sum_{s\in S}\langle x, x_s\rangle x_s$ converges to x.

If $(x_s)_{s\in S}$ is an orthonormal basis in Hilbert space $\mathcal{H}$ then for any $x \in \mathcal{H}$ we have

$$\|x\| = \left(\sum_{s\in S} |\langle x, x_s\rangle|^2\right)^{1/2}.$$

Suppose that $(X_s)_{s\in S}$ is a system of closed linear subspaces of $\mathcal{H}$ which are pairwise orthogonal. If 0 is the only vector from $\mathcal{H}$ which is orthogonal to all X_s then each vector $x \in \mathcal{H}$ can be written as $x = \sum_{s\in S} x_s$ with $x_s \in X_s$. If we have two orthogonal subspaces X_1 and X_2 in a Hilbert space $\mathcal{H}$, then by $X_1 \bigoplus X_2$ we denote direct sum of X_1 and X_2, i.e. the subspace of $\mathcal{H}$ consisting of all vectors $x_1 + x_2$ with $x_i \in X_i$.

1.2 Fourier Transform

Functional series have a long history that can be traced back to the early nineteenth century. French mathematician (and politician) Jean-Baptiste-Joseph Fourier, decomposed a continuous, periodic on $[-\pi, \pi]$ function $f(x)$ into the series od sines and cosines,

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx + b_n \sin nx,$$

where the coefficients a_n and b_n are defined as

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \; dx, \;\; n = 0, 1, 2, \dots \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \; dx, \;\; n = 1, 2, \dots \; . \end{aligned}$$

The sequences $\{a_n, \; n = 0, 1, \dots\}$ and $\{b_n, \; n = 1, 2, \dots\}$ can be viewed as a transform of the original function f. It is interesting that at the time of Fourier's discovery the very notion of function was not precisely defined. Fourier methods have long history in statistics especially in the theory of nonparametric function and density estimation and characteristic functions.

Definition 2 *The Fourier transformation of a function $f \in \mathbb{L}_1(\mathbb{R})$ is defined by*

$$\hat{f}(\omega) = \mathcal{F}[f(x)] = \langle f(x), e^{i\omega x} \rangle = \int_{\mathbb{R}} f(x) \overline{e^{i\omega x}} dx = \int_{\mathbb{R}} f(x) e^{-i\omega x} dx.$$

if $\hat{f} \in \mathbb{L}_1(\mathbb{R})$ is the Fourier transformation of $f \in \mathbb{L}_1(\mathbb{R})$, then

$$f(x) = \mathcal{F}^{-1}[\hat{f}(\omega)] = \frac{1}{2\pi} \int \hat{f}(\omega) e^{i\omega x} d\omega,$$

at every continuity point of f.

The function $\hat{f}(\omega)$ is, in general, a complex function of the form $\hat{f}(\omega) = |\hat{f}(\omega)| e^{i\varphi(\omega)}$. The part $|\hat{f}(\omega)|$ is called the *magnitude spectrum* and the exponent $\varphi(\omega)$ is called the *phase spectrum*.
If $f(x)$ is real, then

- $\hat{f}(-\omega) = \overline{\hat{f}(\omega)}$
- $|\hat{f}(\omega)|$ is an even function and $\varphi(\omega)$ is an odd function of ω.

1.2.1 Basic Properties of the Fourier Transform

Boundedness. $\hat{f} \in \mathbb{L}_\infty(\mathbb{R}),\ \|\hat{f}\|_\infty \leq \|f\|_1$.

Uniform Continuity. $\hat{f}(\omega)$ is uniformly continuous on $-\infty < \omega < \infty$.

Decay. For $f \in \mathbb{L}_1$, $\hat{f}(\omega) \to 0$, when $|\omega| \to \infty$, (Riemann-Lebesgue lemma).

Linearity. $\mathcal{F}[\alpha f(x) + \beta g(x)] = \alpha\mathcal{F}[f(x)] + \beta\mathcal{F}[g(x)]$.

Derivative. $\mathcal{F}[f^{(n)}(x)] = (i\omega)^n \hat{f}(\omega)$.

Plancherel's Identity. $\langle f, g\rangle = \frac{1}{2\pi}\langle \hat{f}, \hat{g}\rangle$; If $g = f$ one obtains Plancherel's identity: $\|f\|^2 = \frac{1}{2\pi}\|\hat{f}\|^2$.

Shifting. $\mathcal{F}[f(x - x_0)] = e^{-i\omega x_0}\hat{f}(\omega)$.

Scaling. $\mathcal{F}[f(ax)] = \frac{1}{|a|}\hat{f}\left(\frac{\omega}{a}\right)$.

Symmetry $\mathcal{F}[\mathcal{F}[f(x)]] = 2\pi f(-x)$.

Convolution. The convolution of f and g is defined as $f * g(x) = \int f(x - t)g(t)dt$. One of the most important properties of Fourier transformation is $\mathcal{F}[f * g(x)] = \hat{f}(\omega)\hat{g}(\omega)$.

Modulation Theorem. From the symmetry property it follows that $f(x)g(x) = \frac{1}{2\pi}F(\omega) * G(\omega)$.

Moment Theorem.

$$\int_{\mathbb{R}} x^n f(x)dx = (i)^n \left.\frac{d^n \hat{f}(\omega)}{d\omega^n}\right|_{\omega=0}.$$

1.2.2 Discrete Fourier Transform

The discrete Fourier transform (DFT) of a sequence $\mathbf{f} = \{f_n,\ n = 0, 1, \ldots, N-1\}$ is defined as

$$\mathbf{F} = \left\{ \sum_{n=0}^{N-1} f_n w_N^{nk},\ k = 0, \ldots, N-1 \right\},$$

where $w_N = e^{-i2\pi/N}$. The inverse is

$$\mathbf{f} = \left\{ \frac{1}{N} \sum_{k=0}^{N-1} F_k w_N^{-nk},\ n = 0, \ldots, N-1 \right\}.$$

The DFT can be interpreted as the multiplication of the input vector by a matrix; therefore, the discrete Fourier transform is a linear operator. If $Q = \{Q_{nk} = e^{-i2\pi nk}\}_{N \times N}$, then $\mathbf{F} = Q \cdot \mathbf{f}$. The matrix Q is unitary (up to a scale factor), i.e., $Q^*Q = NI$, where I is the identity matrix and Q^* is the conjugate transpose of Q.

There are many uses of discrete Fourier transform in statistics. It turns cyclic convolutions into component-wise multiplication, and the fast version of DFT has a low computational complexity of $O(n \log(n))$, meaning that the number of operations needed to transform an input of size n is proportional to $n \log(n)$. For a theory and various other uses of DFT in various fields reader is directed to Brigham [10].

1.3 Wavelets

Wavelet theory has developed into a methodology that is used in a range of disciplines, including mathematics, physics, geophysics, astronomy, signal processing, statistics, and a number of applied fields. Wavelets provide a rich source of already indispensable and intriguing tools for "time-scale" applications. The success of wavelets is attributed to their low computational complexity, good locality and adaptivity, and potential to incorporate prior information about the phenomena. Hence, wavelets are natural tools in modeling complex data structures and multiscale phenomena considered in this project. Wavelet-based methods have also proved to be very advantageous for application to various theoretical statistical problems such as regression, probability density estimation or inverse problems.

Wavelets and wavelet-like decompositions are well suited for analysis of non-stationary and non-isotropic phenomena. They are capable of "zooming-in" and exploring local features at various scales of interest. An additional feature of multiscale methods is that they are "friendly" towards large data sets. In fact, fast filtering algorithms needed to perform wavelet transform exceed in speed classical fast fourier transforms (FFT) and have a calculational complexity of $O(n)$.

1.3.1 A Case Study

We start first with a statistical application of wavelet transforms. This example emphasizes specificity of wavelet-based denoising not shared by standard state-of-art denoising techniques.

A researcher in geology was interested in predicting earthquakes by the level of water in nearby wells. She had a large ($8192 = 2^{13}$ measurements) data set of water levels taken every hour in a period of time of about one year in a California well. Here is the description of the problem.

The ability of water wells to act as strain meters has been observed for centuries. The Chinese, for example, have records of water flowing from wells prior to earthquakes. Lab studies indicate that a seismic slip occurs along a fault prior to rupture. Recent work has attempted to quantify this response, in an effort to use water wells as sensitive indicators of volumetric strain. If this is possible, water wells could aid in earthquake prediction by sensing precursory earthquake strain.

We have water level records from six wells in southern California, collected over a six year time span. At least 13 moderate size earthquakes (magnitude 4.0 - 6.0) occurred in close proximity to the wells during this time interval. There is a a significant amount of noise in the water level record which must first be filtered out. Environmental factors such as earth tides and atmospheric pressure create noise with frequencies ranging from seasonal to semidiurnal. The amount of rainfall also affects the water level, as do surface loading, pumping, recharge (such as an increase in water level due to irrigation), and sonic booms, to name a few. Once the noise is subtracted from the signal, the record can be analyzed for changes in water level, either an increase or a decrease depending upon whether the aquifer is experiencing a tensile or compressional volume strain, just prior to an earthquake.

A plot of the raw data for hourly measurements over one year ($8192 = 2^{13}$ observations) is given in Figure 1a, with a close-up in panel b. After applying the wavelet transform and further processing the wavelet coefficients (thresholding), we obtained a fairly clean signal with a big jump at the earthquake time. The wavelet-denoised data are given in Figure 1d. The magnitude of the water level change at the earthquake time did not get distorted in contrast to traditional smoothing techniques. This local adaptivity is a desirable feature of wavelet methods.

For example, Figure 1c, is denoised signal after applying `supsmo` smoothing procedure. Note that the earthquake jump is smoothed, as well.

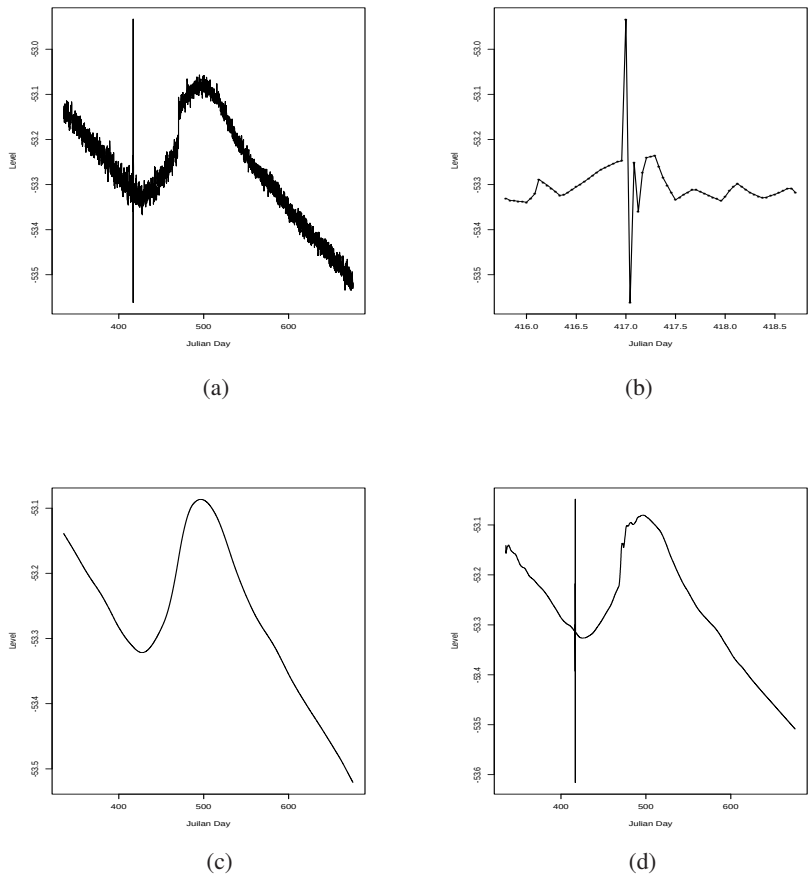


Figure 1: Panel (a) shows $n = 8192$ hourly measurements of the water level for a well in an earthquake zone. Notice the wide range of water levels at the time of an earthquake around $t = 417$. Panel (b) focusses on the data around the earthquake time. Panel (c) demonstrates action of a standard smoother `supsmo`, and (d) gives a wavelet based reconstruction.

1.3.2 Multiresolution Analysis

Fundamental for construction of critically sampled orthogonal wavelets is a notion of multiresolution analysis introduced by Mallat ([45], [46]) A multiresolution analysis (MRA) is a sequence of closed subspaces $V_n, n \in \mathbb{Z}$ in $\mathbb{L}_2(\mathbb{R})$ such that they lie in a containment hierarchy

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots . \tag{1}$$

The nested spaces have an intersection that contains only the zero function and a union that contains all square integrable functions.

$$\cap_n V_j = \{\mathbf{0}\}, \quad \overline{\cup_j V_j} = \mathbb{L}_2(\mathbb{R}).$$

[With $\overline{A}$ we denoted the closure of a set A]. The hierarchy (1) is constructed such that V-spaces are self-similar,

$$f(2^j x) \in V_j \text{ iff } f(x) \in V_0. \tag{2}$$

with the requirement that there exists *a scaling function* $\phi \in V_0$ whose integer-translates span the space V_0,

$$V_0 = \left\{ f \in \mathbb{L}_2(\mathbb{R}) |\ \ f(x) = \sum_k c_k \phi(x-k) \right\},$$

and for which the family $\{\phi(\bullet - k),\ k \in \mathbb{Z}\}$ is an orthonormal basis. It can be assumed that $\int \phi(x)dx \geq 0$. With this assumption this integral is in fact equal to 1. Because of containment $V_0 \subset V_1$, the function $\phi(x) \in V_0$ can be represented as a linear combination of functions from V_1, i.e.,

$$\phi(x) = \sum_{k \in \mathbb{Z}} h_k \sqrt{2}\phi(2x-k), \tag{3}$$

for some coefficients $h_k,\ k \in \mathbb{Z}$. This equation called the *scaling equation* (or two-scale equation) is fundamental in constructing, exploring, and utilizing wavelets.

Theorem 1 *For the scaling function it holds*

$$\int_{\mathbb{R}} \phi(x)dx = 1,$$

or, equivalently,

$$\Phi(0) = 1,$$

where $\Phi(\omega)$ is Fourier transform of ϕ, $\int_{\mathbb{R}} \phi(x)e^{-i\omega x}dx$.

The coefficients h_n in (3) are important in efficient application of wavelet transforms. The (possibly infinite) vector $\mathbf{h} = \{h_n,\ n \in \mathbb{Z}\}$ will be called a *wavelet filter.* It is a low-pass (averaging) filter as will become clear later by its analysis in the Fourier domain.

To further explore properties of multiresolution analysis subspaces and their bases, we will often work in the Fourier domain.

It will be convenient to use Fourier domain for subsequent analysis of wavelet paradigm. Define the function m_0 as follows:

$$m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k \in \mathbb{Z}} h_k e^{-ik\omega} = \frac{1}{\sqrt{2}} H(\omega). \tag{4}$$

The function in (4) is sometimes called the *transfer function* and it describes the behavior of the associated filter $\mathbf{h}$ in the Fourier domain. Notice that the function m_0 is 2π-periodic and that filter taps $\{h_n,\ n \in \mathbb{Z}\}$ are in fact the Fourier coefficients in the Fourier serias of $H(\omega) = \sqrt{2}\, m_0(\omega)$.

In the Fourier domain, the relation (3) becomes

$$\Phi(\omega) = m_0\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right), \tag{5}$$

where $\Phi(\omega)$ is the Fourier transform of $\phi(x)$. Indeed,

$$\begin{aligned}
\Phi(\omega) &= \int_{-\infty}^{\infty} \phi(x) e^{-i\omega x} dx \\
&= \sum_k \sqrt{2}\, h_k \int_{-\infty}^{\infty} \phi(2x-k) e^{-i\omega x} dx \\
&= \sum_k \frac{h_k}{\sqrt{2}} e^{-ik\omega/2} \int_{-\infty}^{\infty} \phi(2x-k) e^{-i(2x-k)\omega/2} d(2x-k) \\
&= \sum_k \frac{h_k}{\sqrt{2}} e^{-ik\omega/2}\ \Phi\left(\frac{\omega}{2}\right) \\
&= m_0\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right).
\end{aligned}$$

By iterating (5), one gets

$$\Phi(\omega) = \prod_{n=1}^{\infty} m_0\left(\frac{\omega}{2^n}\right), \tag{6}$$

which is convergent under very mild conditions concerning the rates of decay of the scaling function ϕ.

Next, we prove two important properties of wavelet filters associated with an orthogonal multiresolution analysis, *normalization* and *orthogonality*.

Normalization.

$$\sum_{k\in\mathbb{Z}} h_k = \sqrt{2}. \tag{7}$$

Proof:

$$\begin{aligned}
\int \phi(x)dx &= \sqrt{2}\sum_k h_k \int \phi(2x-k)dx \\
&= \sqrt{2}\sum_k h_k \,\frac{1}{2}\int \phi(2x-k)d(2x-k) \\
&= \frac{\sqrt{2}}{2}\sum_k h_k \int \phi(x)dx.
\end{aligned}$$

Since $\int \phi(x)dx \neq 0$ by assumption, (7) follows.

This result also follows from $m_0(0) = 1$.

Orthogonality. For any $l \in \mathbb{Z}$,

$$\sum_k h_k h_{k-2l} = \delta_l. \tag{8}$$

Proof: Notice first that from the scaling equation (3) it follows that

$$\begin{aligned}
\phi(x)\phi(x-l) &= \sqrt{2}\,\sum_k h_k \phi(2x-k)\phi(x-l) \\
&= \sqrt{2}\,\sum_k h_k \phi(2x-k)\,\sqrt{2}\,\sum_m h_m \phi(2(x-l)-m).
\end{aligned} \tag{9}$$

By integrating the both sides in (9) we obtain

$$\begin{aligned}\delta_l &= 2\sum_k h_k \left[\sum_m h_m \frac{1}{2}\int \phi(2x-k)\phi(2x-2l-m)\,d(2x)\right] \\ &= \sum_k\sum_m h_k h_m \delta_{k,2l+m} \\ &= \sum_k h_k h_{k-2l}.\end{aligned}$$

The last line is obtained by taking $k = 2l + m$.

An important special case is $l = 0$ for which (8) becomes

$$\sum_k h_k^2 = 1. \tag{10}$$

The fact that the system $\{\phi(\bullet - k),\ k \in \mathbb{Z}\}$ constitutes an orthonormal basis for V_0 can be expressed in the Fourier domain in terms of either $\Phi(\omega)$ or $m_0(\omega)$.

In terms of $\Phi(\omega)$,

$$\sum_{l=-\infty}^{\infty} |\Phi(\omega + 2\pi l)|^2 = 1. \tag{11}$$

From the Plancherel identity and the 2π-periodicity of $e^{i\omega k}$ it follows

$$\begin{aligned}\delta_k &= \int_{\mathbb{R}} \phi(x)\overline{\phi(x-k)}dx \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \Phi(\omega)\overline{\Phi(\omega)}e^{i\omega k}d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{l=-\infty}^{\infty} |\Phi(\omega+2\pi l)|^2 e^{i\omega k}d\omega. \end{aligned}\tag{12}$$

The last line in (12) is the Fourier coefficient a_k in the Fourier series decomposition of

$$f(\omega) = \sum_{l=-\infty}^{\infty} |\Phi(\omega + 2\pi l)|^2.$$

Due to the uniqueness of Fourier representation, $f(\omega) = 1$. As a side result, we obtain that $\Phi(2\pi n) = 0, n \neq 0$, and $\sum_n \phi(x-n) = 1$. The last result follows from inspection of coefficients c_k in the Fourier decomposition of $\sum_n \phi(x-n)$, the series $\sum_k c_k e^{2\pi i k x}$. As this function is

1-periodic,

$$c_k = \int_0^1 \left(\sum_n \phi(x-n) \right) e^{-2\pi i k x} dx = \int_{-\infty}^{\infty} \phi(x) e^{-2\pi i k x} dx = \Phi(2\pi k) = \delta_{0,k}.$$

Remark 1 *Utilizing the identity (11), any set of independent functions spanning V_0, $\{\phi(x-k), k \in \mathbb{Z}\}$, can be orthogonalized in the Fourier domain. The orthonormal basis is generated by integer-shifts of the function*

$$\mathcal{F}^{-1} \left[\frac{\Phi(\omega)}{\sqrt{\sum_{l=-\infty}^{\infty} |\Phi(\omega + 2\pi l)|^2}} \right]. \tag{13}$$

This normalization in the Fourier domain is used in constructing of some wavelet bases.

(**b**) In terms of m_0 :

$$|m_0(\omega)|^2 + |m_0(\omega + \pi)|^2 = 1. \tag{14}$$

Since $\sum_{l=-\infty}^{\infty} |\Phi(2\omega + 2l\pi)|^2 = 1$, then by (5)

$$\sum_{l=-\infty}^{\infty} |m_0(\omega + l\pi)|^2 |\Phi(\omega + l\pi)|^2 = 1. \tag{15}$$

Now split the sum in (15) into two sums – one with odd and the other with even indices, i.e.,

$$\begin{aligned} 1 &= \sum_{k=-\infty}^{\infty} |m_0(\omega + 2k\pi)|^2 |\Phi(\omega + 2k\pi)|^2 + \\ &\quad \sum_{k=-\infty}^{\infty} |m_0(\omega + (2k+1)\pi)|^2 |\Phi(\omega + (2k+1)\pi)|^2. \end{aligned}$$

To simplify the above expression, we use relation (11) and the 2π-periodicity of $m_0(\omega)$.

$$\begin{aligned} 1 &= |m_0(\omega)|^2 \sum_{k=-\infty}^{\infty} |\Phi(\omega + 2k\pi)|^2 + |m_0(\omega + \pi)|^2 \sum_{k=-\infty}^{\infty} |\Phi((\omega + \pi) + 2k\pi)|^2 \\ &= |m_0(\omega)|^2 + |m_0(\omega + \pi)|^2. \end{aligned}$$

Whenever a sequence of subspaces satisfies MRA properties, there exists (though not unique) an orthonormal basis for $\mathbb{L}_2(\mathbb{R})$,

$$\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j,k \in \mathbb{Z}\} \tag{16}$$

such that $\{\psi_{jk}(x),\ j\text{-fixed},\ k \in \mathbb{Z}\}$ is an orthonormal basis of the "difference space" $W_j = V_{j+1} \ominus V_j$. The function $\psi(x) = \psi_{00}(x)$ is called a *wavelet function* or informally *the mother wavelet*.

Next, we discuss the derivation of a wavelet function from the scaling function. Since $\psi(x) \in V_1$ (because of the containment $W_0 \subset V_1$), it can be represented as

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k\ \sqrt{2}\phi(2x-k), \tag{17}$$

for some coefficients $g_k,\ k \in \mathbb{Z}$.

Define

$$m_1(\omega) = \frac{1}{\sqrt{2}}\ \sum_k g_k e^{-ik\omega}. \tag{18}$$

By mimicking what was done with m_0, we obtain the Fourier counterpart of (17),

$$\Psi(\omega) = m_1(\frac{\omega}{2})\Phi(\frac{\omega}{2}). \tag{19}$$

The spaces W_0 and V_0 are orthogonal by construction. Therefore,

$$\begin{aligned} 0 = \int \psi(x)\phi(x-k)dx &= \frac{1}{2\pi}\int \Psi(\omega)\overline{\Phi(\omega)}e^{i\omega k}d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{l=-\infty}^{\infty}\Psi(\omega+2l\pi)\overline{\Phi(\omega+2l\pi)}e^{i\omega k}d\omega. \end{aligned}$$

By repeating the Fourier series argument, as in (11), we conclude

$$\sum_{l=-\infty}^{\infty} \Psi(\omega + 2l\pi)\overline{\Phi(\omega + 2l\pi)} = 0.$$

By taking into account the definitions of m_0 and m_1, and by the derivation as in (14), we find

$$m_1(\omega)\overline{m_0(\omega)} + m_1(\omega + \pi)\overline{m_0(\omega + \pi)} = 0. \tag{20}$$

From (20), we conclude that there exists a function $\lambda(\omega)$ such that

$$(m_1(\omega),\ m_1(\omega + \pi)\ ) = \lambda(\omega)\left(\overline{m_0(\omega + \pi)},\ -\overline{m_0(\omega)}\right). \tag{21}$$

By substituting $\xi = \omega + \pi$ and by using the 2π-periodicity of m_0 and m_1, we conclude that

$$\begin{aligned} \lambda(\omega) &= -\lambda(\omega + \pi), \text{ and} \\ \lambda(\omega) &\quad \text{is } 2\pi\text{-periodic.} \end{aligned} \tag{22}$$

Any function $\lambda(\omega)$ of the form $e^{\pm i\omega}S(2\omega)$, where S is an $\mathbb{L}_2([0, 2\pi])$, 2π-periodic function, will satisfy (20); however, only the functions for which $|\lambda(\omega)| = 1$ will define an orthogonal basis ψ_{jk} of $\mathbb{L}_2(\mathbb{R})$.

To summarize, we choose $\lambda(\omega)$ such that

(i) $\lambda(\omega)$ is 2π-periodic,

(ii) $\lambda(\omega) = -\lambda(\omega + \pi)$, and

(iii) $|\lambda(\omega)|^2 = 1$.

Standard choices for $\lambda(\omega)$ are $-e^{-i\omega}$, $e^{-i\omega}$, and $e^{i\omega}$; however, any other function satisfying (i)-(iii) will generate a valid m_1. We choose to define $m_1(\omega)$ as

$$m_1(\omega) = -e^{-i\omega}\overline{m_0(\omega + \pi)}. \tag{23}$$

since it leads to a convenient and standard connection between the filters $\mathbf{h}$ and $\mathbf{g}$.

The form of m_1 and the equation (11) imply that $\{\psi(\bullet - k),\ k \in \mathbb{Z}\}$ is an orthonormal basis for W_0.

Since $|m_1(\omega)| = |m_0(\omega + \pi)|$, the orthogonality condition (14) can be rewritten as

$$|m_0(\omega)|^2 + |m_1(\omega)|^2 = 1. \tag{24}$$

By comparing the definition of m_1 in (18) with

$$\begin{aligned} m_1(\omega) &= -e^{-i\omega}\frac{1}{\sqrt{2}}\sum_k h_k e^{i(\omega+\pi)k} \\ &= \frac{1}{\sqrt{2}}\sum_k (-1)^{1-k} h_k e^{-i\omega(1-k)} \\ &= \frac{1}{\sqrt{2}}\sum_n (-1)^n h_{1-n} e^{-i\omega n}, \end{aligned}$$

we relate g_n and h_n as

$$g_n = (-1)^n\ h_{1-n}. \tag{25}$$

In signal processing literature, the relation (25) is known as the *quadrature mirror relation* and the filters $\mathbf{h}$ and $\mathbf{g}$ as *quadrature mirror filters.*

Remark 2 *Choosing $\lambda(\omega) = e^{i\omega}$ leads to the rarely used high-pass filter $g_n = (-1)^{n-1}\ h_{-1-n}$. It is convenient to define g_n as $(-1)^n h_{1-n+M}$, where M is a "shift constant." Such re-indexing of* $\mathbf{g}$ *affects only the shift-location of the wavelet function.*

1.3.3 Haar Wavelets

In addition to their simplicity and formidable applicability, Haar wavelets have tremendous educational value. Here we illustrate some of the relations discussed in the Section 1.3.2 using the Haar wavelet. We start with scaling function $\phi(x) = \mathbf{1}(0 \le x \le 1)$ and pretend that everything else is

unknown. By inspection of simple graphs of two scaled Haar wavelets $\phi(2x)$ and $\phi(2x+1)$ stuck to each other, we conclude that the scaling equation (3) is

$$\begin{aligned} \phi(x) &= \phi(2x) + \phi(2x-1) \\ &= \frac{1}{\sqrt{2}}\sqrt{2}\phi(2x) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2x-1), \end{aligned} \tag{26}$$

which yields the wavelet filter coefficients:

$$h_0 = h_1 = \frac{1}{\sqrt{2}}.$$

The transfer functions are

$$m_0(\omega) = \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}e^{-i\omega 0}\right) + \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}e^{-i\omega 1}\right) = \frac{1+e^{-i\omega}}{2}.$$

and

$$m_1(\omega) = -e^{-i\omega}\,\overline{m_0(\omega+\pi)} = -e^{-i\omega}\left(\frac{1}{2} - \frac{1}{2}e^{i\omega}\right) = \frac{1-e^{-i\omega}}{2}.$$

Notice that $m_0(\omega) = |m_0(\omega)|e^{i\varphi(\omega)} = \cos\frac{\omega}{2}\cdot e^{-i\omega/2}$ (after $\cos x = \frac{e^{ix}+e^{-ix}}{2}$). Since $\varphi(\omega) = -\frac{\omega}{2}$, the Haar wavelet has *linear phase,* i.e., the scaling function is symmetric in the time domain. The orthogonality condition $|m_0(\omega)|^2 + |m_1(\omega)|^2 = 1$ is easily verified, as well.

Relation (19) becomes

$$\Psi(\omega) = \frac{1-e^{-i\omega/2}}{2}\Phi\left(\frac{\omega}{2}\right) = \frac{1}{2}\Phi\left(\frac{\omega}{2}\right) - \frac{1}{2}\Phi\left(\frac{\omega}{2}\right)e^{-i\omega/2},$$

and by applying the inverse Fourier transform we obtain

$$\psi(x) = \phi(2x) - \phi(2x-1)$$

in the time-domain. Therefore we "have found" the Haar wavelet function ψ. From the expression for m_1 or by inspecting the representation of $\psi(x)$ by $\phi(2x)$ and $\phi(2x-1)$, we "conclude" that $g_0 = -g_{-1} = \frac{1}{\sqrt{2}}$.

Although the Haar wavelets are well localized in the time domain, in the frequency domain they decay at the slow rate of $O(\frac{1}{n})$ and are not effective in approximating smooth functions.

1.3.4 Daubechies' Wavelets

The most important family of wavelets was discovered by Ingrid Daubechies and fully described in Daubechies [17]. This family is compactly supported with various degrees of smoothness.

The formal derivation of Daubechies' wavelets goes beyond the scope of this chapter, but the filter coefficients of some of its family members can be found by following considerations.

For example, to derive the filter taps of a wavelet with N vanishing moments, or equivalently, $2N$ filter taps, we use the following equations.

The normalization property of scaling function implies

$$\sum_{i=0}^{2N-1} h_i = \sqrt{2},$$

requirement for vanishing moments for wavelet function ψ leads to

$$\sum_{i=0}^{2N-1} (-1)^i i^k h_i = 0,\ k = 0, 1, \ldots, N-1,$$

and, finally, the orthogonality property can be expressed as

$$\sum_{i=0}^{2N-1} h_i h_{i+2k} = \delta_k\ k = 0, 1, \ldots, N-1.$$

We obtained $2N+1$ equations with $2N$ unknowns; however the system is solvable since the equations are not linearly independent.

Example 1 *For $N = 2$, we obtain the system:*

$$\begin{cases} h_0 + h_1 + h_2 + h_3 = \sqrt{2} \\ h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \\ -h_1 + 2h_2 - 3h_3 = 0 \\ h_0\, h_2 + h_1\, h_3 = 0 \end{cases},$$

which has a solution $h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}$, *and* $h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$.

For $N = 4$, *the system is*

$$\begin{cases} h_0 + h_1 + h_2 + h_3 + h_4 + h_5 + h_6 + h_7 = \sqrt{2} \\ h_0^2 + h_1^2 + h_2^2 + h_3^2 + h_4^2 + h_5^2 + h_6^2 + h_7^2 = 1 \\ h_0 - h_1 + h_2 - h_3 + h_4 - h_5 + h_6 - h_7 = 0 \\ h_0h_2 + h_1h_3 + h_2h_4 + h_3h_5 + h_4h_6 + h_5h_7 = 0 \\ h_0h_4 + h_1h_5 + h_2h_6 + h_3h_7 = 0 \\ h_0h_6 + h_1h_7 = 0 \\ 0h_0 - 1h_1 + 2h_2 - 3h_3 + 4h_4 - 5h_5 + 6h_6 - 7h_7 = 0 \\ 0h_0 - 1h_1 + 4h_2 - 9h_3 + 16h_4 - 25h_5 + 36h_6 - 49h_7 = 0 \\ 0h_0 - 1h_1 + 8h_2 - 27h_3 + 64h_4 - 125h_5 + 216h_6 - 343h_7 = 0. \end{cases}$$

Figure 2 depicts two scaling function and wavelet pairs from the Daubechies family. Panels (a) and (b) depict the pair with two vanishing moments, while panels (c) and (d) depict the pair with four vanishing moments.

1.3.5 Discrete Wavelet Transforms

Discrete wavelet transforms (DWT) are applied to discrete data sets and produce discrete outputs. Transforming signals and data vectors by DWT is a process that resembles the fast Fourier transform (FFT), the Fourier method applied to a set of discrete measurements.

Table 1: The analogy between Fourier and wavelet methods

Fourier Methods	Fourier Integrals	Fourier Series	Discrete Fourier Transforms
Wavelet Methods	Continuous Wavelet Transforms	Wavelet Series	Discrete Wavelet Transforms

The analogy between Fourier and wavelet methods is even more complete (Table 1) when we take into account the continuous wavelet transform and wavelet series expansions.

Discrete wavelet transforms map data from the time domain (the original or input data vector) to the wavelet domain. The result is a vector of the same size. Wavelet transforms are linear and they can be defined by matrices of dimension $n \times n$ if they are applied to inputs of size n. Depending

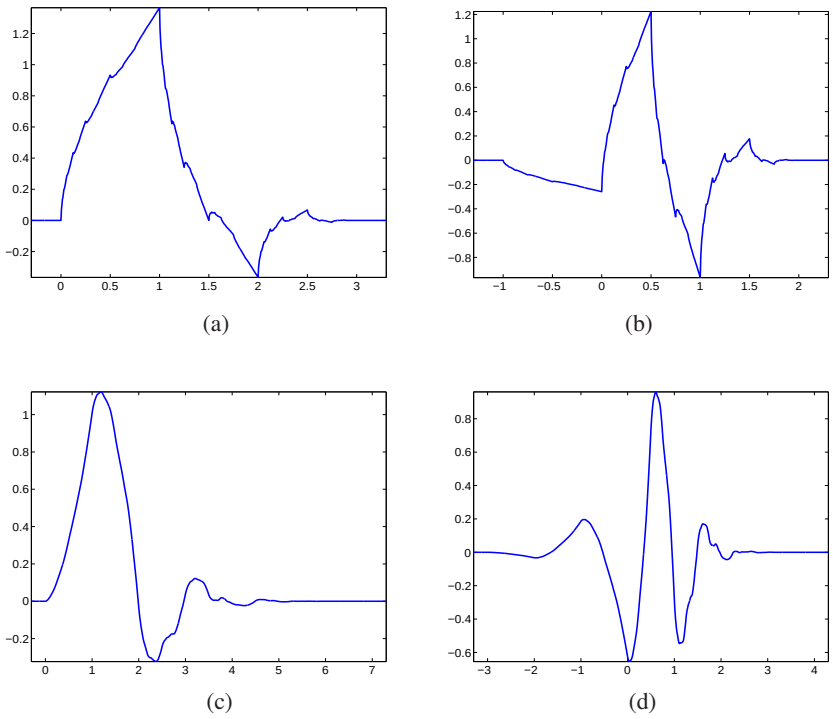


Figure 2: Wavelet functions from Daubechies family. (a) Daubechies scaling function, 2 vanishing moments, 4 tap filter (b) Wavelet function corresponding to (a),(c) Daubechies scaling function, 4 vanishing moments, 8 tap filter (d) Wavelet function corresponding to (c)

on boundary conditions, such matrices can be either orthogonal or "close" to orthogonal. When the matrix is orthogonal, the corresponding transform is a rotation in $\mathbb{R}^n$ in which the data (a n-typle) is a point in $\mathbb{R}^n$. The coordinates of the point in the rotated space comprise the discrete wavelet transform of the original coordinates. Here we provide two toy examples.

Example 2 *Let the vector be* $(1,2)$ *and let* $M(1,2)$ *be the point in* $\mathbb{R}^2$ *with coordinates given by the data vector. The rotation of the coordinate axes by an angle of* $\frac{\pi}{4}$ *can be interpreted as a DWT in the Haar wavelet basis. The rotation matrix is*

$$W = \begin{pmatrix} \cos\frac{\pi}{4} & \sin\frac{\pi}{4} \\ \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

and the discrete wavelet transform of $(1,2)'$ *is* $W \cdot (1,2)' = (\frac{3}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$. *Notice that the energy (squared distance of the point from the origin) is preserved,* $1^2 + 2^2 = (\frac{1}{2})^2 + (\frac{\sqrt{3}}{2})^2$, *since* W *is a*

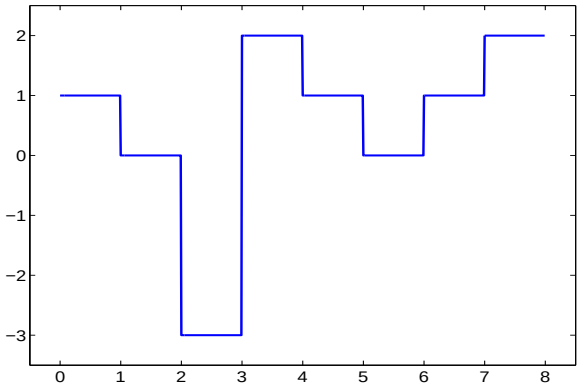


Figure 3: A function interpolating $\mathbf{y}$ on [0,8).

rotation.

Example 3 *Let* $\mathbf{y} = (1, 0, -3, 2, 1, 0, 1, 2)$. *The associated function* f *is given in Figure 3. The values* $f(n) = y_n,\ n = 0, 1, \ldots, 7$ *are interpolated by a piecewise constant function. We assume that* f *belongs to Haar's multiresolution space* V_0.

The following matrix equation gives the connection between $\mathbf{y}$ *and the wavelet coefficients (data in the wavelet domain).*

$$\begin{bmatrix} 1 \\ 0 \\ -3 \\ 2 \\ 1 \\ 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \\ d_{20} \\ d_{21} \\ d_{22} \\ d_{23} \end{bmatrix}.$$

The solution is

$$\begin{bmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \\ d_{20} \\ d_{21} \\ d_{22} \\ d_{23} \end{bmatrix} = \begin{bmatrix} \sqrt{2} \\ -\sqrt{2} \\ 1 \\ -1 \\ \frac{1}{\sqrt{2}} \\ -\frac{5}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Thus,

$$\begin{aligned} f &= \sqrt{2}\phi_{-3,0} - \sqrt{2}\psi_{-3,0} + \psi_{-2,0} - \psi_{-2,1} \\ &\quad + \frac{1}{\sqrt{2}}\psi_{-1,0} - \frac{5}{\sqrt{2}}\psi_{-1,1} + \frac{1}{\sqrt{2}}\psi_{-1,2} - \frac{1}{\sqrt{2}}\psi_{-1,3}. \end{aligned} \tag{27}$$

The solution is easy to verify. For example, when $x \in [0, 1)$,

$$f(x) = \sqrt{2} \cdot \frac{1}{2\sqrt{2}} - \sqrt{2} \cdot \frac{1}{2\sqrt{2}} + 1 \cdot \frac{1}{2} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = 1/2 + 1/2 = 1 \ (= y_0).$$

Applying wavelet transforms by multiplying the input vector with an appropriate orthogonal matrix is conceptually straightforward task, but of limited practical value. Storing and manipulating the transformation matrices for long inputs ($n > 2000$) may not even be feasible.

This obstacle is solved by the link of discrete wavelet transforms with fast filtering algorithms from the field of signal and image processing.

1.3.6 Nondecimated Wavelet Transform

An important development in the statistical context has been the routine use of the nondecimated wavelet transform (NDWT), also called the stationary or translation-invariant wavelet transform. See, for example Nason and Silverman [55].

The NDWT is obtained by modifying the Mallat DWT algorithm: at each stage, no decimation takes place but instead the filters are padded out with alternate zeroes to double their length. The

effect (roughly speaking, depending on boundary conditions) is to yield an overdetermined transform with n coefficients at each of $\log 2n$ levels. The transform contains the standard DWT for every possible choice of time origin. Since the NDWT is no longer (1–1), it does not have a unique inverse, but the DWT algorithm is easily modified to yield the average basis inverse (Coifman and Donoho [15]), which gives the average of the DWT reconstructions over all choices of time origin. Both the NDWT and the average basis reconstruction are $O(n \log n)$ algorithms.

1.3.7 Directional Wavelet Transforms

Definition 3 *A wavelet ψ is said to be directional if and only if its Fourier transform is strictly supported in a convex cone with apex at the origin.*

Example 4 (Cauchy Wavelets) *Let $\mathcal{C}$ be the cone defined by the angles α and β and let $\tilde{\alpha} = \alpha + \pi/2, \tilde{\beta} = \beta - \pi/2$. The Cauchy wavelet is defined by*

$$\widehat{\psi}^{C}_{l,m}(\vec{\omega}) = \begin{cases} \left(\vec{\omega} \cdot \vec{1}_{\tilde{\alpha}}\right)^{l} \left(\vec{\omega} \cdot \vec{1}_{\tilde{\beta}}\right)^{m} e^{-\vec{\omega}\cdot\vec{\eta}}, & \text{if } \vec{\omega} \in C \\ 0, & \text{otherwise} \end{cases}$$

where $l, m \in \mathbb{N}$, $\vec{1}_{\alpha}$ stands for the unit vector in the direction α and $\vec{\eta}$ is a vector in the cone defined by the angles $\tilde{\alpha}$ and $\tilde{\beta}$

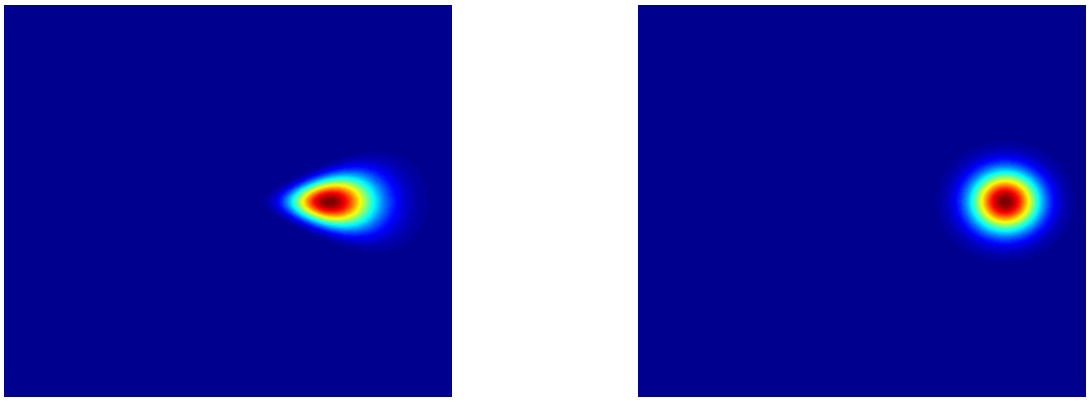
Example 5 (Gabor or Morlet Wavelets) *Let $\vec{\omega}_0$ be a fixed frequency vector. The 2-D Morlet wavelet is defined by*

$$\widehat{\psi}(\vec{\omega}) = e^{-\|\vec{\omega}-\vec{\omega}_0\|^2}.$$

For more information and reference on directional wavelets we direct the reader to Vandergheynst [71].

1.3.8 Daubechies–Lagarias Algorithm

A challenge in exhibiting a wavelet-based classificator is computational. Namely, except for the Haar wavelet, all compactly supported orthonormal families of wavelets (e.g., Daubechies, Symmlet, Coiflet, etc.) scaling and wavelet functions have no closed form. A non-elegant solution is to have values of the mother and father wavelet given in a table. Evaluation of $\phi_{jk}(x)$ or $\psi_{jk}(x)$, for given x, then can be performed by interpolating the table values.



(a) (b)

Figure 4: a) The Cauchy wavelet in frequency plane, with support in the cone $C = C(-\pi/6, \pi/6)$. b) The Morlet wavelet in frequency plane.

Based on Daubechies and Lagarias ([18], [19]) *local pyramidal algorithm* a solution is proposed. A brief theoretical description and MATLAB program are provided.

Let ϕ be the scaling function of a compactly supported wavelet generating an orthogonal MRA. Suppose the support of ϕ is $[0, N]$. Let $x \in (0, 1)$, and let $dyad(x) = \{d_1, d_2, \ldots, d_n, \ldots\}$ be the set of 0-1 digits in dyadic representation of x $(x = \sum_{j=1}^{\infty} d_j 2^{-j})$. By $dyad(x, n)$ we denote the subset of the first n digits from $dyad(x)$, i.e., $dyad(x, n) = \{d_1, d_2, \ldots, d_n\}$.

Let $h = (h_0, h_1, \ldots, h_N)$ be the vector of wavelet filter coefficients. Define two $N \times N$ matrices as

$$T_0 = \sqrt{2}(h_{2i-j-1})_{1 \le i,j \le N}, \quad \text{and} \quad T_1 = \sqrt{2}(h_{2i-j})_{1 \le i,j \le N}. \tag{28}$$

Then

Theorem 2 *(Daubechies and Lagarias)*

$$\lim_{n \to \infty} T_{d_1} \cdot T_{d_2} \cdot \cdots \cdot T_{d_n} = \begin{bmatrix} \phi(x) & \phi(x) & \ldots & \phi(x) \\ \phi(x+1) & \phi(x+1) & \ldots & \phi(x+1) \\ \vdots & & & \\ \phi(x+N-1) & \phi(x+N-1) & \ldots & \phi(x+N-1) \end{bmatrix}. \tag{29}$$

The convergence of $||T_{d_1} \cdot T_{d_2} \cdots \cdots T_{d_n} - T_{d_1} \cdot T_{d_2} \cdots \cdots T_{d_{n+m}}||$ to zero, for fixed m, is exponential and constructive, i.e., effective bounds, that decrease exponentially to 0, can be established.

Example: Consider the DAUB 2 wavelet basis ($N = 3$). The corresponding filter is ($\frac{1+\sqrt{3}}{4\sqrt{2}}$, $\frac{3-\sqrt{3}}{4\sqrt{2}}$, $\frac{3+\sqrt{3}}{4\sqrt{2}}$, $\frac{1-\sqrt{3}}{4\sqrt{2}}$). According to (28) the matrices T_0 and T_1 are given as:

$$T_0 = \begin{bmatrix} \frac{1+\sqrt{3}}{4} & 0 & 0 \\ \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} \\ 0 & \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} \end{bmatrix}, \text{ and } \quad T_1 = \begin{bmatrix} \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} & 0 \\ \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4} \\ 0 & 0 & \frac{1-\sqrt{3}}{4} \end{bmatrix}.$$

If, for instance, $x = 0.45$, then $dyad(0.45, 20) = \{$ 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1 $\}$. The values $\phi(0.45), \phi(1.45)$, and $\phi(2.45)$ are calculated as

$$\prod_{i \in dyad(0.45,20)} T_i = \begin{bmatrix} 0.86480582 & 0.86480459 & 0.86480336 \\ 0.08641418 & 0.08641568 & 0.08641719 \\ 0.04878000 & 0.04877973 & 0.04877945 \end{bmatrix}.$$

By using the so-called two-scale equations, it is possible to give an algorithm for calculating values of the mother wavelet, the function ψ_{jk}, see Vidakovic [73]. For our purposes direct calculation of wavelet coefficients is unnecessary since, having scaling coefficients at some level J, all wavelet coefficients at coarser levels can be obtained utilizing Mallat's algorithm. MATLAB codes are available in the Appendix.

1.3.9 Wavelets in Statistics

Wavelets found application in a remarkable diversity of disciplines: mathematics, physics, numerical analysis, signal and image processing, and many others. In this section we will briefly describe some of the application of wavelets in statistics.

Denoising/Shrinkage Consider the following model

$$y_i = f_i + \epsilon_i \quad i = 1, \ldots, n,$$

where y_i is the observed data, f_i is the unknown signal and ϵ is the random noise. Wavelet shrinkage usually refers to the reconstruction of the unknown signal f obtained from the

shrunk wavelet coefficients. Shrinking and truncating the data directly or the coefficients in their Fourier series expansions is an old technique in signal and image processing. For non-local bases, such as trigonometric, shrinking the coefficients can affect the global shape of the reconstructed function and introduce unwanted artifact. In the context of function estimation by wavelets, the shrinkage has an additional feature; it is connected with smoothing (denosing) because the measures of smoothness of a function depend on the magnitudes of its wavelet coefficients.

The two most common thresholding methods are *hard* and *soft*. The analytic expression for the hard and soft thresholding rules are

$$\delta^h(d,\lambda) = d\mathbf{1}(|d| > \lambda) \quad \lambda \geq 0, d \in \mathbb{R},$$

and

$$\begin{aligned}\delta^s(d,\lambda) &= (d - \text{sign}(d)\cdot\lambda)\mathbf{1}(|d| > \lambda) \\ &= \text{sign}(d)(|d| - \lambda)_+ \quad \lambda \geq 0, d \in \mathbb{R},\end{aligned}$$

respectively. The rules are illustrated in Figure 5.

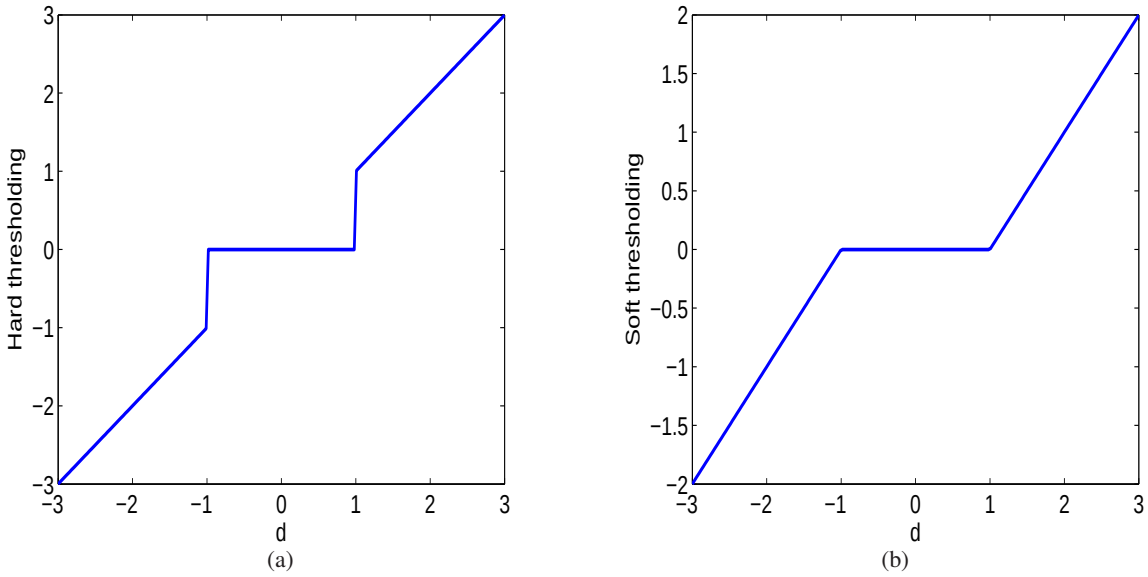


Figure 5: a) Hard and b) Soft thresholding rules for $\lambda = 1$.

One of the simples choices of threshold is the *universal* threshold, proposed by Donoho and

Johnstone [22]

$$\lambda_{un} = \sigma\sqrt{2\log(n)}/\sqrt{n}$$

with σ replaced by a suitable estimate $\hat{\sigma}$ derived from the data when σ is unknown.

Literature is rich with various shrinkage methods based on wavelets. Brief overview of the various methods can be found in Chapter 6 in Vidakovic [73]. In Chapter 3 of this thesis we propose two new methods of shrinkage, based on the testing of multiple hypothesis.

Density estimation Given the realization $X_1, X_2, \ldots, X_n$ of a random variable X with an unknown density function f, it is of interest to estimate f. An automatic approximation to f is the empirical "density" $f_e(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - X_i)$. It is easy to see that f_e is an unbiased estimator of f since $\mathbb{E}f_e(x) = \delta * f = f$. However, for an absolutely continuous underling distribution, the estimator f_e is a poor choice. It is not smooth, moreover, it is even not a function. For these reasons, different estimators have been proposed. The density estimation has a long history and many solutions. The local nature of wavelets functions promises superiority over projection estimators that use classical orthonormal bases (Fourier, Haar, etc.) The wavelet estimators are simple, well-localized in space/frequency, and share a variety of optimality properties. We refer reader to Chapter 7 in Vidakovic [73], for detailed information on various method of density estimation using wavelets.

Inverse problems (See Abramovich, Bailey and Sapatinas [1]) Some interesting scientific applications involve indirect noisy measurements. For example, the primary interest might be in a function g, but the data are only accessible from some linear transformation Kg and, in addition, are corrupted by noise. In this case, the estimation of g from indirect noisy observations $\mathbf{y} = (y_1, \ldots, y_n)'$ is often referred to in statistics as a *linear inverse problem*. Such linear inverse problems arise in a wide variety of scientific settings with different types of transformations K. Examples include applications in estimating financial derivatives, in medical imaging(the Radon transform), in magnetic resonance imaging (the Fourier transform) and in spectroscopy (convolutional transformations). Typically such problems are referred to as *ill posed* when naive estimate of g obtained from the inverse transform K^{-1} applied to an estimate of Kg fails to produce reasonable results because K^{-1} is an unbounded linear operator

and the presence of even small amount of noise in the data 'blow up' when the straightforward inversion estimate is used. Wavelets and wavelet-like decomposition provide solution for the inverse problem. For more information please refer to the work of Abramovich and Silverman [4]

Changepoint problems (See Abramovich, Bailey and Sapatinas [1]) The good time-frequency localization of wavelets provides a natural motivation for their use in changepoint problems. Here the main goal is the estimation of the number, locations and size of a function's abrupt changes, such as sharp spikes or jumps. Changepoint models are used in a wide set of practical problems in quality control, medicine, economics and physical sciences. The detection of edges and the location of sharp contrast in digital pictures in signal processing and image analysis is also fall within the general changepoint problem framework (see, for example, Mallat and Hwang [48] and Ogden [59]).

The general idea used to detect a function's abrupt changes through a wavelet approach is based on the connection between the function's local regularity properties at a certain point and the rate of the decay of the wavelet coefficients near this point across increasing resolution levels (see, for example Daubechies [17] section 2.9, and Mallat and Hwang [48]). Local singularities are identified by 'unusual' behavior in the wavelet coefficient at high resolution levels at the corresponding locations. Such ideas are discussed in more detail in, for example, Wang [75] and Raimondo [65]. Bailey et al. [9] have provided an example of related work in detecting transient underwater sound signals.

Antoniadis and Gijbels [7] also proposed an 'indirect' wavelet-based method for detecting and locating changepoints before curve fitting. Conventional function estimation techniques are then used on each identified segment to recover overall curve. They have shown that, provided that discontinuities can be detected and located with sufficient accuracy, detection followed by wavelet smoothing enjoys optimal rates of convergence.

Bayesian approaches using wavelets have also been suggested for estimation the locations and magnitudes of a function's changepoints and are discussed by Richwine [66] and Ogden and Lynch [60]. These place a prior distribution on a changepoint and then examine the posterior

distribution of the changepoint given the estimated wavelet coefficients.

Time series analysis (See, Vidakovic [73] Chapter 9 and Abramovich, Bailey and Sapatinas [1])
The chief motivation for modeling a time series is the need for forecasting. To that tend, an analysis of time series can be performed in the time domain as well as in the frequency domain. Wavelets provide additional insight in the analysis of time series via scale analysis. The notion of frequency in Fourier analysis can be related to the notion of scale in multiscale analysis; and often Fourier-based tools for exploring time series have their wavelet counterparts (for instance, wavelet spectra, wavelet periodogram, and scalogram). Self-similarity properties of some processes, such as fractional Brownian motion or ARIMA, can be well described by wavelet methods.

- *Spectral density estimation for stationary processes.* Consider a real-valued, stationary, Gaussian random process $Y_1, Y_2, \ldots$ with zero mean and covariance function $R(j) = \mathrm{cov}(Y_k, Y_{k+1})$. An important tool in the analysis of such process is its spectral density (or power spectrum function) $f(\omega)$, which is the Fourier transform of the covariance function $R(j)$:

$$f(\omega) = R(0) + 2\sum_{j=0}^{\infty} R(j)\cos(2\pi j\omega). \tag{30}$$

  Given n observations $y_1, \ldots, y_n$, the sample estimator of the spectral density is the sample spectrum or *periodogram* $I(\omega)$. This is essentially the square of the discrete Fourier transform of the data and it is typically computed at the 'fundamental' frequencies $\omega_j = j/n, j = 0, \ldots, n/2$, in which case

$$I(\omega_j) = \frac{1}{n}\left|\sum_{k=1}^{n} \exp -2i\pi(k-1)\omega_j\right|^2, \quad \omega_j = j/n,\ j = 0, \ldots, n/2.$$

  $I(\omega)$ is an asymptotically unbiased estimate of $f(\omega)$, but it is not a consistent estimate and it is usually suggested that the periodogram should be smoothed to estimate spectral density.

  The conventional smoothing is based on kernel estimation, but wavelet-based techniques can also be used for estimating $\log(f(\omega))$ from $\log(I(\omega))$. Moulin [53] and Gao [30]

proposed wavelet shrinkage procedures for estimating $\log(f(\omega))$, and then exponentiating to obtain an estimate of $f(\omega)$. Please see Vidakovic [73] Chapter 9 for more details.

- *Wavelet Analysis of non-stationary processes.* The conventional theory of spectral analysis discussed above applies only to stationary processes. The spectral characteristics of non-stationary processes change over time, so these precesses do not have a spectral density as defined through equation (30). Instead, it is natural to consider the evolution of their spectral properties in time. Owning to their localization in both the time and the frequency, or scale or resolution domains, wavelets are a natural choice for this purpose. Recently, there has been growing interest in wavelets analysis of non-stationary processes, particularly in so-called 'locally stationary' processes (see, for example, Neumann [57], von Sachs and Shneider [68], Neumann and von Sachs [58], and Nason et al. [56]).

  By analogy with the periodogram which is the square of the discrete Fourier transform of the data, we can define a *wavelet periodogram* as the square of the DWT of the data (in fact, it is preferable to define the wavelet periodogram via the non-decimated wavelet transform (NDWT) of the data). The wavelet periodogram may then be used as a tool for analyzing how the spectra characteristics of a process change in time. As in the case of stationary time series, the wavelet periodogram need to be denoised to obtain a consistent estimate, and that is achieved through appropriate thresholding procedures. For a discussion and more details we refer to Nason and von Sachs [54]. More general coverage of wavelet methods in the analysis of non-stationary time series mayu be found in Priestley [64], Morttin [52] or Percival and Walden [63].

There are many other application of wavelets in statistics. For more information on application of wavelets in statistics, please refer to Vidakovic [73], Antoniadis and Oppenheim [8], and Abramovich, Bailey and Sapatinas [1].

CHAPTER II

WAVELET-BASED GENERALIZED LINEAR CLASSIFIER

2.1 Chapter Introduction

Classification is one of the main statistical procedures in the field of Pattern Recognition Theory. Based on historic (training) covariate measurements (univariate or multivariate) the decision-maker is to classify a newly obtained observation. For instance, an observation may be classified as conforming or non-conforming, low or high, real or fake, black or white, etc, depending on the problem context. This unknown nature of the observation will be called a *class,* and in this paper we consider problems possessing only two possible exclusive classes, "0" and "1." Formally, the classifier is a function that maps the d-dimensional space of covariates to the set $\{0, 1\}$.

In this chapter we are concerned with classifier functions represented by wavelet decompositions. Our proposal builds on the existing theory of generalized linear classifiers by Devroye, Györfi, and Lugosi [21]. Kohler [41] argues that the use of standard wavelets in the general regression may produce suboptimal results if the distribution of the design is non-uniform. It is likely that the same holds true for wavelet-based classifiers. However, we have found that in practical and simulated situations, when design distribution is clearly non-uniform, our classificator works well.

2.2 The Bayes Classification Problem

In this section we introduce the Bayes classification problem.

Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ denote a random variable. Let μ be the probability measure of X and η the regression of Y on X, i.e., for a Borel set $A \in \mathbb{R}^d$

$$\mu(A) = P\{X \in A\}$$

and

$$\eta(x) = P\{Y = 1|X = x\} = \mathbb{E}\{Y|X = x\}.$$

It can be demonstrated that the pair (μ, η) uniquely determines the distribution of (X, Y).

Any function $g : \mathbb{R}^d \to \{0, 1\}$ is a classifier. For a classifier g, the error (risk) function is the probability of error, i.e., $L(g) = P\{g(X) \neq Y\}$.

The Bayes classifier

$$g^*(x) = \mathbf{1}(\eta(x) > 1/2)$$

minimizes L, i.e., for any classifier g,

$$P\{g^*(X) = Y\} \geq P\{g(X) = Y\},$$

or equivalently, $L(g^*) \leq L(g)$.

We will denote this minimal error with L^* and call it Bayes error . The attribute Bayes comes from the fact that classification is made according to the posterior probability,

$$\eta(X) = P\{Y = 1|X\}.$$

Let $D_n = ((X_1, Y_1) \ldots, (X_n, Y_n))$ be a training set and X be a new observation. A classifier constructed on the basis of D_n is denoted by g_n. Label Y is classified by the decision $g_n(X) = g_n(X, D_n)$. The error probability is

$$L_n = L_n(g_n) = P\{Y \neq g_n(X)|D_n\}.$$

For more details and results about general Bayes classification problem we direct the reader to the excellent monograph by Devroye, Györfi, and Lugosi [21].

2.2.1 Consistency

One of the desirable properties of any estimation and classification rule is its consistency. If a training data is given $D_n = ((X_1, Y_1) \ldots, (X_n, Y_n))$, the best one can expect from a classifier is to achieve the Bayes error probability L^*. Generally, one cannot hope to obtain a classifier that exactly achieves L^*, but it is possible to construct a sequence of classifiers g_n, such that the error probability

$$L_n = L(g_n) = P\{g_n(X, D_n) \neq Y|D_n\}$$

gets arbitrarily close to L^* with large probability (that is, for "most" D_n). This idea is formulated in the definition of *consistency*:

Definition 4 (Weak and Strong Consistency) *A classification rule is consistent (or asymptotically Bayes-risk efficient) for a certain distribution of* (X, Y) *if*

$$\mathbb{E}L_n = P\{g_n(X, D_n) \neq Y\} \to L^* \quad as \quad n \to \infty,$$

and strongly consistent if

$$\lim_{n\to\infty} L_n = L^* \quad with\ probability\ 1.$$

"A consistent rule guarantees that by increasing the amount of data the probability that the error probability is within a very small distance of the optimal achievable gets arbitrary close to one. Intuitively, the rule can eventually learn the optimal decision from a large amount of training data with high probability. Strong consistency means that by using more data the error probability gets arbitrary close to the optimum for every training sequence except for a set of sequences that has zero probability altogether.

A decision rule can be consistent for a certain class of distributions of (X, Y), but may not be consistent for others. It is clearly desirable to have a rule that is consistent for a large class of distributions. Since in many situations we do not have any prior information about the distribution, it is essential to have a rule that gives good performance for all distributions." Devroye, Györfi, and Lugosi [21].

This strong requirement is formulated in the following definition:

Definition 5 *A sequence of decision rules is called universally (strongly) consistent if it is (strongly) consistent for any distribution of the pair* (X, Y).

2.2.2 Previous Work

This work closely follows the result of Devroye, Györfi, and Lugosi [21] (Theorem A. 1) and the work of Chang, Kim, and Vidakovic [13]. Devroye, Györfi, and Lugosi [21] defined the generalized linear classifier g_n by

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{j=1}^{k_n} a_j^* \psi_j(x) \leq 0 \\ 1 & \text{otherwise.} \end{cases}$$

where coefficients $a_1^*, \ldots, a_{k_n}^*$ minimize the empirical squared error

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{k_n} a_j\psi_j(X_i) - (2Y_i - 1)\right)^2$$

under certain the constraints. The sequence ψ_j forms an uniformly bounded ($|\psi_j(x)| \leq 1$ for all x and j) orthonormal basis. Devroye, Györfi, and Lugosi [21] propose that under some regularity conditions the classifier defined above will be universally consistent in both strong and weak senses. Most series classifiers have unattractive feature, they are not local in nature – points at arbitrary distances from x affect the decision at x. In this work we propose a generalized linear classifier which is based on wavelet sequence. Wavelets are local in nature, but this comes at the price. It is well known that wavelets are not uniformly bounded, this will change minimization constraints and consistency conditions. Section 2.3 defines the wavelet-based generalized linear classifier and states consistency results.

Now, we will briefly introduce the classifier proposed in Chang, Kim, and Vidakovic [13].

Let's assume that a density of X, $f \in \mathbb{L}_2(\mathbb{R})$, exists. If $f_0 \in \mathbb{L}_2(\mathbb{R})$ and $f_1 \in \mathbb{L}_2(\mathbb{R})$ are class-conditional densities, i.e., densities of X when $Y = 0$ and $Y = 1$ respectively, and p and $1-p$ class probabilities, $P(Y = 1)$ and $P(Y = 0)$, then the function

$$\alpha(x) = pf_1(x) - (1-p)f_0(x)$$

has the representation $(2\eta(x) - 1)f(x)$, and the classifier g^* can be written as

$$g^*(x) = \mathbf{1}(\alpha(x) > 0). \tag{31}$$

Since $\alpha(x)$ is in L_2, it allows the wavelet representation

$$\alpha(x) = \sum_{k\in Z} c_{J,k}\phi_{J,k}(x) + \sum_{j\geq J}\sum_{k\in Z} d_{j,k}\psi_{j,k}(x).$$

A *raw* wavelet-based linear classifier, $\hat{g}_J$, is defined as

$$\hat{g}_J(x) = \mathbf{1}(\hat{\alpha}_J(x) > 0), \tag{32}$$

where $\hat{\alpha}_J(x)$ is an estimator of the projection of α on V_J, i.e. an estimator of $\alpha_J(x) = \sum_{k\in Z} c_{J,k}\phi_{J,k}(x)$.

The coefficients $c_{J,k} = \int_R (2\eta(x) - 1) f(x) \phi_{J,k}(x)\, dx = E[(2\eta(X) - 1)\phi_{J,k}(X)]$ can be, by moment matching, estimated by $\hat{c}^n_{J,k} = \frac{1}{n} \sum_{i=1}^n (2Y_i - 1)\phi_{J,k}(X_i)$. Thus, one can take $\hat{\alpha}_{n,J}(x) = \sum_k \hat{c}^n_{J,k} \phi_{J,k}(x)$, and the estimator from (32) can be rewritten as $\hat{g}_{n,J}(x) = \mathbf{1}(\hat{\alpha}_{n,J}(x) > 0)$. If the wavelet basis is interpolating, or close to interpolating, then the coefficients $\{\hat{c}^n_{J,k},\ k \in Z\}$ can be thought as values of α sampled at equally spaced points. Let $\hat{L}_n(J) = P(Y \neq \hat{g}_{n,J}(X) | D_n)$ be the error probability of $\hat{g}_{n,J}$.

Results of Chang, Kim, and Vidakovic [13] show that the estimator $\hat{g}_{n,J}(x)$ is weakly consistent. They also establish what the linear estimator $\hat{g}_{n,J}$ gains in performance if regularized. Regularization is achieved by wavelet shrinkage. Soft shrinkage with universal threshold is used. The regularized estimator for the training sample of size n, multiresolution level J, and threshold λ is denoted by $\tilde{g}_{n,J,\lambda}$. The classifier $\tilde{g}_{n,J,\lambda}$ is also consistent in the weak sense.

Generalized linear classifier presented here is of a different nature; it is distribution free, and based on the minimization of the empirical squared error. Classifier proposed by Chang, Kim, and Vidakovic [13], is not universal, since they need to assume the existence of square integrable conditional densities. It is important to develop a distribution free methods, since, after all, the distribution of (X, Y) is unknown in practice. Besides, even if we have an iid sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ at out disposal, we do not know of any test verifying whether conditional densities of X are square integrable. The next section defines the wavelet-based generalized linear classifier and states the result about its consistency.

2.3 Wavelet Based Classifier

The wavelet based classifier is preceded in the literature by the Fourier series classifier. All such classifiers can be put in the form: classify $X = x$ to be in class 0 if $\sum_{j=1}^k a_{n,j} \psi_j(x) \leq 0$. Functions ψ_j are fixed and represent the basis for the series estimate. Coefficients $a_{n,j}$ depend on the training sample of size n. The number of basis functions is denoted by k and usually regulates smoothness.

The literature on Fourier series classifiers is rich. Work by Van Ryzin [70], Greblicki and his team (Greblicki [31]; Greblicki and Rutkovski [32]; Greblicki and Pawlak ([33], [34]) explore various theoretical concepts of consistency and rates of convergence of the classifiers.

Any $\mathbb{L}_2(\mathbb{R})$ function f can be represented as

$$f(x) = \sum_{j,k} d_{jk}\psi_{jk}(x),$$

and this unique representation corresponds to a multiresolution decomposition $\mathbb{L}_2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} W_j$. Also, for any fixed J the decomposition $\mathbb{L}_2(\mathbb{R}) = V_J \bigoplus_{j=J}^{\infty} W_j$ corresponds to the representation

$$f(x) = \sum_k c_{J,k}\phi_{Jk}(x) + \sum_{j>J}\sum_k d_{jk}\psi_{jk}(x),$$

where

$$c_{J,k} = \int f(x)\phi_{Jk}(x)dx, \quad d_{jk} = \int f(x)\psi_{jk}(x)dx.$$

Consider the wavelet-based generalized linear classifier of the form

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{j,k\in K_n} d_{jk}\psi_{jk}(x) \le 0 \\ 1 & \text{otherwise,} \end{cases}$$

where $\psi_{jk} \in \mathbb{L}_2(\mu)$, $j,k \in \mathbb{Z}$ are fixed wavelet sequence, and the coefficients d_{jk} are estimated from the training sequence D_n by minimizing the empirical squared error

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j,k\in K_n} d_{jk}\psi_{jk}(X_i) - (2Y_i - 1)\right)^2.$$

The following result shows that the above classifier is universally consistent.

Theorem 3 *Let $\psi_{jk} \in \mathbb{L}_2(\mu)$, $j,k \in \mathbb{Z}$ be a wavelet sequence, such that $|\psi(x)| \le W$ for some W. Let the coefficients d^*_{jk} minimize the empirical square error*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j,k\in K_n} d_{jk}\psi_{jk}(X_i) - (2Y_i - 1)\right)^2$$

under the constraint $\sum_{j,k\in K_n} |d_{jk}|2^{j/2} \le b_n$, $b_n \ge 1$. Define the generalized linear classifier g_n by

$$g_n(x) = \begin{cases} 0 & \textit{if } \sum_{j,k\in K_n} d^*_{jk}\psi_{jk}(x) \le 0 \\ 1 & \textit{otherwise.} \end{cases}$$

If K_n and b_n satisfy

$$|K_n| \to \infty, \quad b_n \to \infty \quad \text{and} \quad \frac{|K_n| b_n^4 \log(b_n)}{n} \to 0,$$

then $\mathbb{E}\{L(g_n)\} \to L^*$ for all distributions of (X,Y), that is, the rule g_n is universally consistent. Under the same conditions the rule g_n is strongly universally consistent

2.3.1 Proof of the Theorem

Let $\delta > 0$ be arbitrary. Then there exists a constant M such that $P\{\|X\| > M\} < \delta$.

$$\begin{aligned}
L(g_n) - L^* &= P\{g_n(X, D_n) \neq Y | D_n\} - P\{g^*(X) \neq Y\} \\
&= P\{g_n(X, D_n) \neq Y, \|X\| \leq M | D_n\} + P\{g_n(X, D_n) \neq Y, \|X\| > M | D_n\} \\
&\quad - P\{g^*(X) \neq Y, \|X\| \leq M\} - P\{g^*(X) \neq Y, \|X\| > M\} \\
&= P\{g_n(X, D_n) \neq Y, \|X\| > M | D_n\} - P\{g^*(X) \neq Y, \|X\| > M\} \\
&\quad + P\{g_n(X, D_n) \neq Y, \|X\| \leq M | D_n\} - P\{g^*(X) \neq Y, \|X\| \leq M\}
\end{aligned}$$

Note that:

$$\begin{aligned}
&P\{g_n(X, D_n) \neq Y, \|X\| > M | D_n\} - P\{g^*(X) \neq Y, \|X\| > M\} \\
&= P\{\|X\| > M\} \left[P\{g_n(X, D_n) \neq Y | \, \|X\| > M, D_n\} - P\{g^*(X) \neq Y | \, \|X\| > M\}\right] \\
&\leq P\{\|X\| > M\} < \delta
\end{aligned}$$

Thus,

$$L(g_n) - L^* \leq \delta + P\{g_n(X) \neq Y, \|X\| \leq M | D_n\} - P\{g^*(X) \neq Y, \|X\| \leq M\}.$$

It suffice to show that $P\{g_n(X) \neq Y, \|X\| \leq M | D_n\} - P\{g^*(X) \neq Y, \|X\| \leq M\} \to 0$ in the required sense for every $M > 0$. Introduce the notation $f_n^*(x) = \sum_{j,k \in K_n} d_{jk} \psi_{jk}(x)$. By Corollary A. 1 (see Appendix A), we see that

$$\begin{aligned}
&P\{g_n(X) \neq Y, \|X\| \leq M | D_n\} - P\{g^*(X) \neq Y, \|X\| \leq M\} \\
&\leq \sqrt{\int_{\|x\| \leq M} \left(f_n^*(x) - (2\eta(x) - 1)\right)^2 \, \mu(dx)}
\end{aligned}$$

We prove that the right-hand side converges to zero in probability. Observe that since $\mathbb{E}\{2Y - 1 | \, X = x\} = 2\eta(x) - 1$, for any function $h(x)$,

$$\begin{aligned}
&\mathbb{E}\{(h(X) - Y)^2 | \, X = x\} \\
&\quad = \mathbb{E}\{(h(x) - \eta(x) + \eta(x) - Y)^2 | \, X = x\} \\
&\quad = (h(x) - \eta(x))^2 + 2(h(x) - \eta(x))\mathbb{E}\{\eta(x) - Y | \, X = x\} + \mathbb{E}\{(\eta(X) - Y)^2 | \, X = x\} \\
&\quad = (h(x) - \eta(x))^2 + \mathbb{E}\{(\eta(X) - Y)^2 | \, X = x\}
\end{aligned}$$

hence,

$$
\begin{aligned}
&(h(x) - (2\eta(x) - 1))^2 \\
&= \mathbb{E}\{(h(X) - (2Y - 1))^2 | \, X = x\} - \mathbb{E}\{((2Y - 1) - (2\eta(X) - 1))^2 | \, X = x\}
\end{aligned}
$$

therefore, denoting the class of functions over which we minimize by

$$
\mathcal{F}_n = \left\{ \sum_{j,k \in K_n} d_{jk} \psi_{jk} : \sum_{j,k \in K_n} |d_{jk}| 2^{j/2} \leq b_n \right\},
$$

we have

$$
\begin{aligned}
&\int_{\|x\| \leq M} (f_n^*(x) - (2\eta(x) - 1))^2 \, \mu(dx) \\
&\quad = \mathbb{E}\left\{ (f_n^*(X) - (2Y - 1))^2 I_{\|X\| \leq M} | \, D_n \right\} - \mathbb{E}\left\{ ((2Y - 1) - (2\eta(X) - 1))^2 I_{\|X\| \leq M} \right\} \\
&\quad = \left(\mathbb{E}\left\{ (f_n^*(X) - (2Y - 1))^2 I_{\|X\| \leq M} | \, D_n \right\} - \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y - 1))^2 I_{\|X\| \leq M}\} \right) \\
&\quad + \left(\inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y - 1))^2 I_{\|X\| \leq M}\} - \mathbb{E}\left\{ ((2Y - 1) - (2\eta(X) - 1))^2 I_{\|X\| \leq M} \right\} \right).
\end{aligned}
$$

The last two terms may be combined to yield

$$
\inf_{f \in \mathcal{F}_n} \int_{\|x\| \leq M} (f(x) - (2\eta(x) - 1))^2 \, \mu(dx)
$$

which converges to zero since wavelets form an orthonormal basis in $\mathbb{L}_2(\mu)$. To prove that the first term converges to zero in probability, observe that we may assume without loss of generality that $P\{\|X\| \leq M\} = 0$. Now

$$
\begin{aligned}
&\mathbb{E}\left\{ (f_n^*(X) - (2Y - 1))^2 I_{\|X\| \leq M} | \, D_n \right\} - \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y - 1))^2 I_{\|X\| \leq M}\} \\
&\quad = \mathbb{E}\left\{ (f_n^*(X) - (2Y - 1))^2 | \, D_n \right\} - \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y - 1))^2\} \\
&\quad = \mathbb{E}\left\{ (f_n^*(X) - (2Y - 1))^2 | \, D_n \right\} - \frac{1}{n} \sum_{i=1}^{n} (f_n^*(X_i) - (2Y_i - 1))^2 \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} (f_n^*(X_i) - (2Y_i - 1))^2 - \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y - 1))^2\} \leq
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left\{(f_n^*(X)-(2Y-1))^2|\, D_n\right\} - \frac{1}{n}\sum_{i=1}^n (f_n^*(X_i)-(2Y_i-1))^2 \\
&+ \sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-(2Y_i-1))^2 - \mathbb{E}\{(f(X)-(2Y-1))^2\}\right| \\
&\leq 2\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-(2Y_i-1))^2 - \mathbb{E}\{(f(X)-(2Y-1))^2\}\right| \\
&= 2\sup_{h\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \mathbb{E}\{h(X,Y)\}\right|,
\end{aligned}
$$

where the class of functions $\mathcal{T}$ is defined by

$$
\mathcal{T} = \left\{h(x,y) = (f(x)-(2y-1))^2 :\ f\in\mathcal{F}_n\right\}.
$$

Observe that since $|2y-1|=1$ and there exist constant $W>1$ such that $|\psi(x)|\leq W$, we have

$$
\begin{aligned}
&0 \leq h(x,y) = (f(x)-(2y-1))^2 \\
&= \left(\sum_{j,k\in K_n} d_{jk}\psi_{jk}(x) - (2y-1)\right)^2 = \left(\sum_{j,k\in K_n} d_{jk}2^{j/2}\psi(2^jx-k) - (2y-1)\right)^2 \\
&\leq \left(\sum_{j,k\in K_n} |d_{jk}|2^{j/2}|\psi(2^jx-k)| + |2y-1|\right)^2 \leq \left(\sum_{j,k\in K_n} |d_{jk}|2^{j/2}W + 1\right)^2 \\
&\leq 2\left(\left(\sum_{j,k\in K_n} |d_{jk}|2^{j/2}W\right)^2 + 1\right) \leq 2(W^2b_n^2+1) \leq 4W^2b_n^2.
\end{aligned}
$$

Therefore, Theorem A. 2 (see Appendix A) asserts that

$$
\begin{aligned}
&P\left\{\mathbb{E}\left\{(f_n^*(X)-(2Y-1))^2|\, D_n\right\} - \inf_{f\in\mathcal{F}_n}\mathbb{E}\{(f(X)-(2Y-1))^2\} > \epsilon\right\} \\
&\leq P\left\{\sup_{h\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \mathbb{E}\{h(X,Y)\}\right| > \epsilon/2\right\} \\
&\leq 8\mathbb{E}\left\{\mathcal{N}\left(\frac{\epsilon}{16},\mathcal{T}(Z_1^n)\right)\right\} e^{-n\epsilon^2/(512(4W^2b_n^2)^2)},
\end{aligned}
$$

where $Z_1^n = (X_1,Y_1),\ldots,(X_n,Y_n)$.

Next for fixed z_1^n, we estimate covering number $\mathcal{N}\left(\frac{\epsilon}{16},\mathcal{T}(z_1^n)\right)$ (Definition A. 1). For arbitrary $f_1, f_2\in\mathcal{F}_n$, consider the functions $h_1(x,y) = (f_1(x)-(2y-1))^2$ and $h_2(x,y) = (f_2(x)-(2y-$

$1))^2$. Then for any probability measure ν on $\mathbb{R}^d \times \{0,1\}$,

$$\begin{aligned}
&\int |h_1(x,y) - h_2(x,y)|\, \nu(d(x,y)) \\
&\quad = \int |(f_1(x) - (2y-1))^2 - (f_2(x) - (2y-1))^2|\, \nu(d(x,y)) \\
&\quad = \int |f_1^2(x) - f_2^2(x) - 2(2y-1)(f_1(x) - f_2(x))|\, \nu(d(x,y)) \\
&\quad = \int |f_1(x) - f_2(x)||f_1(x) + f_2(x) - 2(2y-1)|\, \nu(d(x,y)) \\
&\quad = \int |f_1(x) - f_2(x)| \left| \sum_{j,k \in K_n} d_{jk}^{(1)} \psi_{jk}(x) + \sum_{j,k \in K_n} d_{jk}^{(2)} \psi_{jk}(x) - 2(2y-1) \right| \nu(d(x,y)) \\
&\quad \le \int |f_1(x) - f_2(x)| \left(\sum_{j,k \in K_n} |d_{jk}^{(1)}| 2^{j/2} |\psi(2^j x - k)| \right. \\
&\qquad\qquad \left. + \sum_{j,k \in K_n} |d_{jk}^{(2)}| 2^{j/2} |\psi(2^j x - k)| + 2 \right) \nu(d(x,y)) \\
&\quad \le \int |f_1(x) - f_2(x)| [2W b_n + 2]\, \nu(d(x,y)) \le 2W(b_n+1) \int |f_1(x) - f_2(x)|\, \mu(dx)
\end{aligned}$$

where μ is the marginal measure for ν on $\mathbb{R}^d$. Thus, for any $z_1^n = (x_1, y_1), \ldots, (x_n, y_n)$ and ϵ,

$$\mathcal{N}(\epsilon, \mathcal{T}(z_1^n)) \le \mathcal{N}\left(\frac{\epsilon}{2W(b_n+1)}, \mathcal{F}(x_1^n) \right).$$

Therefore, it suffices to estimate the covering number corresponding to $\mathcal{F}_n$. Since $\mathcal{F}_n$ is a subset of a linear space of functions, we have $V_{\mathcal{F}_n^+} \le |K_n| + 1$ (see Definitions A. 2, A.3 and Theorem A. 3). By Corollary A. 2 (see Appendix A),

$$\begin{aligned}
&\mathcal{N}\left(\frac{\epsilon}{2W(b_n+1)}, \mathcal{F}_n(x_1^n) \right) \le \left(\frac{4eWb_n}{\epsilon/(2W(b_n+1))} \log \left(\frac{2eWb_n}{\epsilon/(2W(b_n+1))} \right) \right)^{|K_n|+1} \\
&\le \left(\frac{8e^2W^2b_n^2}{\epsilon^2/(4W^2(b_n+1)^2)} \right)^{|K_n|+1} \le \left(\frac{32e^2W^2b_n^2(b_n+1)^2}{\epsilon^2} \right)^{|K_n|+1}.
\end{aligned}$$

Summarizing, we have

$$\begin{aligned}
&P\left\{ \mathbb{E}\left\{ (f_n^*(X) - (2Y-1))^2 \middle| D_n \right\} - \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - (2Y-1))^2\} > \epsilon \right\} \\
&\quad \le 8 \left(\frac{2^{13}e^2W^4b_n^2(b_n+1)^2}{\epsilon^2} \right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)} \\
&\quad \le 8 \left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2} \right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)}
\end{aligned}$$

which goes to zero if $|K_n| b_n^4 \log(b_n)/n \to 0$. This proves that the rule g_n is universally consistent. Strongly universal consistency follows by applying the Borel-Cantelli lemma (Lemma A. 1 and Theorem A. 4) to the last probability.

$$\sum_n P\left\{\mathbb{E}\left\{(f_n^*(X)-(2Y-1))^2|\, D_n\right\} - \inf_{f\in\mathcal{F}_n} \mathbb{E}\{(f(X)-(2Y-1))^2\} > \epsilon\right\}$$
$$\leq 8\sum_n \left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2}\right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)}$$

According to a well known calculus theorem (Cauchy root test (Theorem A. 5.)) if

$$\lim_{n\to\infty}\left(\left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2}\right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)}\right)^{\frac{1}{n}} < 1$$

then

$$\sum_n \left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2}\right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)} < \infty.$$

To verify that

$$\left(\left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2}\right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)}\right)^{\frac{1}{n}}$$
$$= \exp\left\{4\left(\frac{|K_n|+1}{n}\right)\log\left(\frac{2^{15/4}e^{1/2}Wb_n}{\epsilon^{1/2}}\right) - \frac{\epsilon^2}{2^{13}W^4b_n^4}\right\}$$

Now

$$\exp\left\{4\left(\frac{|K_n|+1}{n}\right)\log\left(\frac{2^{15/4}e^{1/2}Wb_n}{\epsilon^{1/2}}\right) - \frac{\epsilon^2}{2^{13}W^4b_n^4}\right\} < 1$$

if and only if

$$4\left(\frac{|K_n|+1}{n}\right)\log\left(\frac{2^{15/4}e^{1/2}Wb_n}{\epsilon^{1/2}}\right) - \frac{\epsilon^2}{2^{13}W^4b_n^4} < 0$$

if and only if

$$\frac{2^{15}W^4(|K_n|+1)b_n^4\log\left(\frac{2^{15/4}e^{1/2}Wb_n}{\epsilon^{1/2}}\right)}{n\epsilon^2} < 1.$$

For every $\epsilon > 0$, under the condition $|K_n| b_n^4 \log(b_n)/n \to 0$,

$$\frac{2^{15}W^4(|K_n|+1)b_n^4\log\left(\frac{2^{15/4}e^{1/2}Wb_n}{\epsilon^{1/2}}\right)}{n\epsilon^2} \to 0$$

therefore

$$\lim_{n\to\infty}\left(\left(\frac{2^{15}e^2W^4b_n^4}{\epsilon^2}\right)^{|K_n|+1} e^{-n\epsilon^2/(2^{13}W^4b_n^4)}\right)^{\frac{1}{n}} < 1.$$

□

2.4 Implementation

In this section we discuss the practical aspects of the wavelet-based classifier. We propose two practical ways of constructing wavelet-based generalized linear classifier.

2.4.1 Single Scale Classifier

We begin with the selection of wavelet function. Obtained results indicate that the selection of the wavelet function might improve the results. Next step is to select a multiresolution level J. Let the scaling function $\phi(x)$ generates orthonormal multiresolution analysis (MRA) and let the multiresolution subspace V_J be spanned by the functions $\{\phi_{J,k}(x) = 2^{J/2}\phi(2^J x - k),\ k \in \mathbb{Z}\}$. Recall, what with the fixed J the decomposition $\mathbb{L}_2(\mathbb{R}) = V_J \bigoplus_{j=J}^{\infty} W_j$ corresponds to the representation

$$f(x) = \sum_k c_{J,k}\phi_{Jk}(x) + \sum_{j>J}\sum_k d_{jk}\psi_{jk}(x),$$

for any $f \in \mathbb{L}_2(\mathbb{R})$. We restrict ourselves only to multiresolution subspace V_J. Now, if X's are rescaled to $[0,1]$ then $K_n = \{0, 1, 2, \ldots 2^J - 1\}$. Wavelet classifier will be of the form

$$g_{n,J}(x) = \begin{cases} 0 & \text{if } \sum_{k=0}^{2^J} d^*_{Jk}\phi_{Jk}(x) \leq 0 \\ 1 & \text{otherwise,} \end{cases}$$

where the coefficients d^*_{Jk} minimize the empirical square error

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{2^J-1} d_{Jk}\phi_{Jk}(X_i) - (2Y_i - 1)\right)^2$$

under the proper constraints. Minimization of the squared error is attractive because there are efficient algorithms to find the minimizing coefficients, while minimizing the number of errors committed on the training sequence is computationally more difficult. Daubechies-Lagarias algorithm was used for evaluation of $\phi_{jk}(x)$ or $\psi_{jk}(x)$, for given x. For details on the algorithm please refer to Chapter I.

As in Chang, Kim, and Vidakovic [13] our experiments indicate that the wavelet-based classifier g_n gains in performance if regularized. Regularization is achieved by wavelet shrinkage; wavelet coefficients d^*_{Jk} are thresholded using BAMS method (see Vidakovic and Ruggeri [74]) with `Symmlet 8` as a wavelet function.

2.4.2 Classifier Based on Multiple Scales

The single scale method described above is not very effective. One of the reasons for that is that by use of single scale, we completely ignore ability of wavelets to capture different features of the signal using different scales. In addition, for a single fixed scale J and location set $K_n = \{0, 1, 2, \dots 2^J - 1\}$ function $\phi_{Jk}(x)$ will be zero equal for many k. Hence, when we perform minimization many of the resulting coefficients d^*_{Jk} will be insignificant, this is one possible explanation of the improvements of the results after regularization. Another consideration is the number of parameters in the model. Fixed single scale J and location set $K_n = \{0, 1, 2, \dots 2^J - 1\}$ will generate 2^J coefficients d^*_{Jk}. Thus, with small sample sizes we a limited in the selection of scale. To overcome these issues, we recommend using the following method based on multiple scales. After the set of scales $\{J_1, \dots J_l\}$ is specified, for each scale $j = J_t$ the location set will be defined $K_j = \{k \in 0 \dots 2^j - 1 : \phi_{jk}(x) \neq 0\}$. Note that X's are rescaled to $[0, 1]$ as in the case above.

Wavelet classifier will be of the form

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{j\in\{J_1,\dots J_l\}} \sum_{k\in K_j} d^*_{jk}\phi_{jk}(x) \leq 0 \\ 1 & \text{otherwise,} \end{cases}$$

where the coefficients d^*_{jk} minimize the empirical square error

$$\frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j\in\{J_1,\dots J_l\}} \sum_{k\in K_j} d_{jk}\phi_{jk}(X_i) - (2Y_i - 1) \right)^2,$$

under the proper constrains. Length of the set K_j will now depend on the length of the filter of the selected wavelet function.

2.5 Examples

We discuss in detail simulational studies in which the true classes are known, and real-life examples from industrial and bioinformatical fields.

The performance of the classifier g_n constructed using training data set of size n and evaluated at testing data $\{(X_j, Y_j) : j = 1, \dots, m\}$, is measured using empirical error

$$\hat{L}(g_n) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{I}(g_n(X_j) \neq Y_j).$$

In all the following examples WBGLC will denote the proposed single scale classifier, $\text{WBGLC}_{\text{reg}}$ its regularized version, and WBGLC_{MS} denotes classifier based on multiple scales.

The first two examples are taken form Chang, Kim, and Vidakovic [13].

2.5.1 Simulated Data Example: 0 - 1 Classification

In this simulation we want to classify between observations coming from two different normal populations.

The training set, $\{(X_i, Y_i), i = 1, \ldots, n\}$, ($n$ is even) is generated as follows. For the first half of the data, $X_i,\ i = 1, \ldots, \frac{n}{2}$ are sampled from the standard normal distribution and $Y_i = 1,\ i = 1, \ldots, \frac{n}{2}$. For the second half, $X_i,\ i = \frac{n}{2} + 1, \ldots, n$ are sampled from normal distribution with mean 2 and variance 1, while $Y_i = 0,\ i = \frac{n}{2} + 1, \ldots, n$.

The validation set $\{(X_j, Y_j), j = 1, \ldots, m\}$ is generated in the same way. We compare the empirical errors WBGLC with $\text{WBGLC}_{\text{reg}}$ and the error of the logistic regression classifier,

$$L_n^{\text{logit}}(m) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{I}\Big(\mathbf{I}(f(X_j) > 0.5) \neq Y_j\Big),$$

where f is fitted logistic regression.

The results for various values of n and $m = 200$ are given in Table 2. Results from Chang, Kim, and Vidakovic [13] for the same data example are given in Table 3 for the purpose of comparison. The empirical error of the regularized estimator from Chang, Kim, and Vidakovic [13] is denoted $\tilde{L}_n(J, m, \lambda)$.

Sample size n	WBGLC	$\text{WBGLC}_{\text{reg}}$	Logistic Regression
80	0.2838	0.2410	0.1589
200	0.2024	0.1886	0.1606
400	0.1756	0.1705	0.1587
800	0.1659	0.1639	0.1587
2000	0.1641	0.1625	0.1585

Table 2: Average of the empirical errors over 1000 simulations using n training data points, $J = 6$ if $N < 1000$, $J = 7$ for $N > 1000$, and $m = 200$ validation data points. Wavelet function – Daubechies 16

n	$\hat{L}_n(6, 200)$	$\tilde{L}_n(6, 200, \lambda)$	Logistic Regression
80	0.272	0.170	0.158
200	0.200	0.164	0.154
400	0.187	0.176	0.171
800	0.169	0.163	0.163
2000	0.160	0.157	0.158

Table 3: Results from Chang, Kim, and Vidakovic [13]. Empirical errors using n training data points, $J = 6$, and $m = 200$ validation data points

2.5.2 Simulated Data Example: 0 - 1 - 0 Classification

In the following simulated example the linear logistic regression classifier is not appropriate, so it is not surprising that logistic regression performs so poorly.

We generate the training data set, $\{(X_i, Y_i), i = 1, \ldots, n\}$, ($n$ is a multiple of 3) as follows. In the first third of the data, $X_i,\ i = 1, \ldots, \frac{n}{3}$ is generated from a normal distribution with mean -2 and variance 1, with $Y_i = 0,\ i = 1, \ldots, \frac{n}{3}$. In the second third of the data, $X_i,\ i = \frac{n}{3} + 1, \ldots, \frac{2n}{3}$ are standard normal random variables and $Y_i = 1,\ i = \frac{n}{3} + 1, \ldots, \frac{2n}{3}$. Finally, in the last third of the data, $X_i,\ i = \frac{2n}{3} + 1, \ldots, n$ are generated from normal distribution with mean 2 and variance 1, and $Y_i = 0,\ i = \frac{2n}{3} + 1, \ldots, n$.

The evaluation set $\{(X_j, Y_j), j = 1, \ldots, m\}$ is generated in an analogous manner. The results for various values of n and $m = 300$ are presented in Table 4. Results from Chang, Kim, and Vidakovic [13] for the same data example are given in Table 5 for the purpose of comparison. The empirical error of the regularized estimator from Chang, Kim, and Vidakovic [13] is denoted by $\tilde{L}_n(J, m, \lambda)$.

We compare the empirical errors of WBGLC with $\text{WBGLC}_{\text{reg}}$ and the error of the logistic regression classifier,

$$L_n^{\text{logit}}(m) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{I}\Big(\mathbf{I}(f(X_j) > 1/3) \neq Y_j\Big),$$

where f is fitted logistic regression.

2.5.3 Two Simple Examples

The next two examples demonstrate experimentally consistency results proved in the Theorem 3. In both examples performance of the classifier based on various sample sized of the training data set

n	WBGLC	$\text{WBGLC}_{\text{reg}}$	Logistic Regression
120	0.2983	0.2626	0.5017
300	0.2434	0.2280	0.5011
600	0.2245	0.2174	0.5008
900	0.2196	0.2152	0.5007
1200	0.2232	0.2162	0.5007

Table 4: Average of the empirical errors over 1000 simulations using n training data points, $J = 6$ if $N < 1000$, $J = 7$ for $N > 1000$, and $m = 300$ evaluation data points. Wavelet function – Daubechies 16.

n	$\hat{L}_n(7, 300)$	$\tilde{L}_n(7, 300, \lambda)$
120	0.340	0.213
300	0.288	0.221
600	0.247	0.218
900	0.232	0.212
1200	0.214	0.202

Table 5: Results from Chang, Kim, and Vidakovic [13]. Average empirical errors using training data of size n, $J = 7$, and $m = 300$ evaluation data points.

is measured on the testing data set of 200 observations.

Example 1. Let

$$Y = \begin{cases} 1 & \text{if} \quad X_1 + X_2 + X_3 < 3 \\ 0 & \text{otherwise.} \end{cases}$$

If X_1, X_2, and X_3 are known, Y is known as well. The Bayes classifier decides 1 if $X_1+X_2+X_3 < 3$ and 0 otherwise. The corresponding Bayes probability of error is zero. But lets assume that X_3 is not available to the observer, and we would only have access to X_1 and X_2. Given X_1 and X_2, when should we guess that $Y = 1$? To answer this question, one must know the joint distribution of (X_1, X_2, X_3), or, equivalently, the joint distribution of (X_1, X_2, Y). So lets assume that X_1, X_2, and X_3 are i.i.d. exponential random variables with mean 1 (i.e. they have density e^{-u} on $[0, \infty)$). The Bayes rule compares $P\{Y = 1|\, X_1, X_2\}$ with $P\{Y = 0|\, X_1, X_2\}$ and makes decision consistent with the maximum of these two values. A simple calculation shows that

$$\begin{aligned} \eta(X_1, X_2) = P\{Y = 1|\, X_1, X_2\} &= P\{X_1 + X_2 + X_3 < 3|\, X_1, X_2\} \\ &= P\{X_3 < 3 - X_1 - X_2|\, X_1, X_2\} \\ &= \max(0, 1 - e^{-(3-X_1-X_2)}). \end{aligned}$$

Thus, the Bayes classifier is

$$g^*(X_1, X_2) = \begin{cases} 1 & \text{if} \quad X_1 + X_2 < 3 - \log 2 \\ 0 & \text{otherwise.} \end{cases}$$

The probability of error is

$$\begin{aligned} L^* =& P\{g^*(X_1, X_2) \neq Y\} = P\{X_1 + X_2 < 3 - \log 2, X_1 + X_2 + X_3 \geq 3\} + \\ & P\{X_1 + X_2 \geq 3 - \log 2, X_1 + X_2 + X_3 < 3\} \\ =& \mathbb{E}\{e^{-(3-X_1-X_2)}\mathbf{I}_{\{X_1+X_2<3-\log 2\}}\} + \\ & \mathbb{E}\{(1 - e^{-(3-X_1-X_2)})\mathbf{I}_{\{3-\log 2 \leq X_1+X_2<3\}}\} \\ =& \int_0^{3-\log 2} xe^{-x}e^{-(3-x)}dx + \int_{3-\log 2}^{3} xe^{-x}(1 - e^{-(3-x)})dx \\ =& 0.1710. \end{aligned}$$

Note, the density of $X_1 + X_2$ is ue^{-u} on $[0, \infty)$.

Now, assume that observer has access only to X_1, then the Bayes classifier is allowed to use X_1 only. Similarly, we have

$$\eta(X_1) = P\{Y = 1 |\ X_1\} = P\{X_3 + X_2 < 3 - X_1 |\ X_1\} = \max(0, 1 - (1 + 3 - X_1)e^{-(3-X_1)}).$$

The cross over at $1/2$ occurs at $X_1 = c = 1.3216$. Thus the Bayes classifier is given by

$$g^*(X_1) = \begin{cases} 1 & \text{if} \quad X_1 < c \\ 0 & \text{otherwise.} \end{cases}$$

The probability of error is

$$\begin{aligned} L^* =& P\{g^*(X_1, X_2) \neq Y\} = P\{X_1 < c, X_1 + X_2 + X_3 \geq 3\} + P\{X_1 \geq c, X_1 + X_2 + X_3 < 3\} \\ =& \mathbb{E}\{(1 + 3 - X_1)e^{-(3-X_1)}\mathbf{I}_{\{X_1<c\}}\} + \\ & \mathbb{E}\{(1 - (1 + 3 - X_1)e^{-(3-X_1)})\mathbf{I}_{\{c \leq X_1<3\}}\} \\ =& \int_0^c e^{-x}(1 + 3 - x)e^{-(3-x)}dx + \int_c^3 e^{-x}(1 - (1 + 3 - x)e^{-(3-x)})dx \\ =& 0.2829. \end{aligned}$$

The Bayes error has increased. Finally, if we do not have access to any of the three variables, the best we can do is see which class is most likely. To this end, we compute

$$P\{Y=0\} = P\{X_1+X_2+X_3 \geq 3\} = (1+3+3^2/2)e^{-3} = 0.4232.$$

If we set $g \equiv 1$ all the time, we make an error with probability 0.4232. Tables 6 and 7 provide simulational results, the confirm theoretical findings, as empirical error converges to theoretical Bayes error as sample size increases. Figure 6 gives graphical illustration of the results of the simulations.

Method	100	250	500	1000
WBGLC	0.2107	0.1839	0.1784	0.1821
$\text{WBGLC}_{\text{reg}}$	0.1970	0.1792	0.1755	0.1763
WBGLC_{MS}	0.2326	0.1894	0.1752	0.1729
Logistic Regression	0.1953	0.1943	0.1962	0.1976

Table 6: Example 1: X_1 and X_2 are observable. Average of the empirical errora of the 100 simulations, for various training sample sizes. Wavelet function is Daubechies with 8 vanishing moments. The Bayes error is 0.1710

Method	100	250	500	1000
WBGLC	0.3407	0.3080	0.3028	0.2910
$\text{WBGLC}_{\text{reg}}$	0.3725	0.3427	0.2910	0.2866
WBGLC_{MS}	0.3561	0.3243	0.3062	0.2904
Logistic Regression	0.3415	0.3107	0.3097	0.3101

Table 7: Example 1: Only X_1 is observable. Average of the empirical errors of the 100 simulations, for various training sample sizes. Wavelet function is Daubechies with 8 vanishing moments. The Bayes error is 0.2829

Example 2. This example is almost identical to the previous example, with the exception of the distribution of X_i.

$$Y = \begin{cases} 1 & \text{if} \quad X_1^2+X_2^2+X_3^2 < c \\ 0 & \text{otherwise,} \end{cases}$$

where X_i are iid $N(0,1)$ random variables and c is a quantile of χ^2_3 (chi squared with three degrees of freedom) distribution.

In our experiments we selected $c = \chi^2_3(0.5) = 2.36597$, this implies that classes have equal probabilities,

$$P\{Y=0\} = P\{X_1^2+X_2^2+X_3^2 \geq c\} = 0.5 = P\{Y=1\}.$$

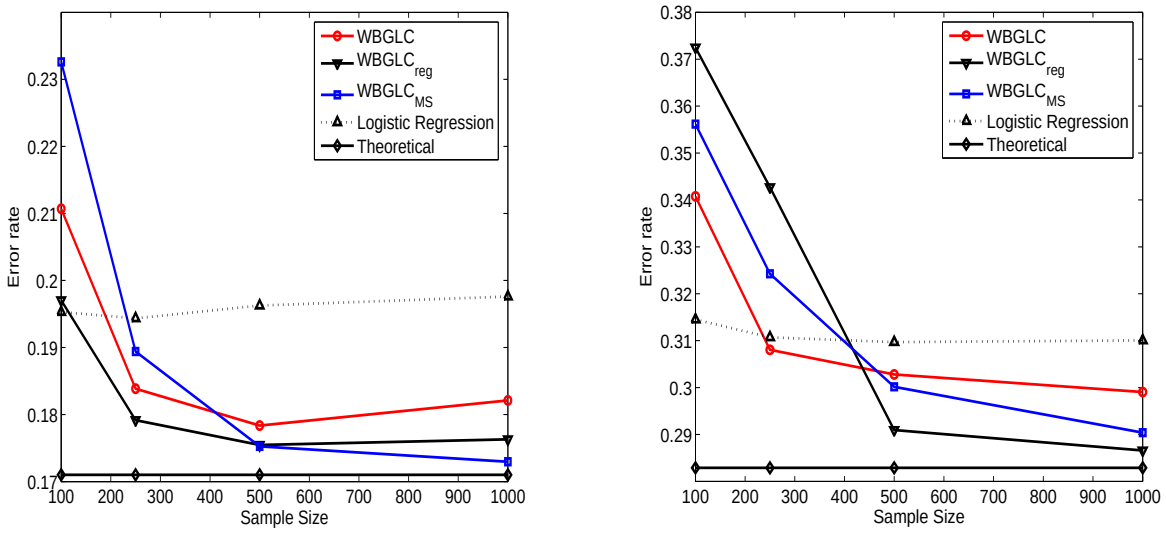


Figure 6: Example 1: Average of the empirical errors of the 100 simulations, for various training sample sizes. (a) Case 1: X_1 and X_2 are observable. (b) Case 2: Only X_1 is observable.

Let's assume that only X_1 and X_2 are observable. Simple calculation shows

$$\begin{aligned}\eta(X_1, X_2) = P\{Y = 1|\ X_1, X_2\} =& P\{X_1^2 + X_2^2 + X_3^2 < c|\ X_1, X_2\} \\ =& P\{X_3^2 < c - X_1^2 - X_2^2|\ X_1, X_2\} \\ =& \max(0, F_1(c - X_1^2 - X_2^2)),\end{aligned}$$

where F_1 denotes the cdf of χ_1^2 distribution. Thus Bayes decision function is

$$g^*(X_1, X_2) = \begin{cases} 1 & \text{if} \quad X_1^2 + X_2^2 < c - F_1^{-1}(1/2) = 1.9110 \\ 0 & \text{otherwise.} \end{cases}$$

The probability of error is

$$\begin{aligned}L^* =& P\{g^*(X_1, X_2) \neq Y\} = P\{X_1^2 + X_2^2 + X_3^2 < c, X_1^2 + X_2^2 > c - F_1^{-1}(1/2)\} + \\ & P\{X_1^2 + X_2^2 + X_3^2 \geq c, X_1^2 + X_2^2 \leq c - F_1^{-1}(1/2)\} \\ =& \int_{c - F_1^{-1}(1/2)}^{c} F_1(c - x) f_2(x) dx + \int_0^{c - F_1^{-1}(1/2)} (1 - F_1(c - x) f_2(x) dx, \\ =& 0.1702\end{aligned}$$

where f_2 denotes pdf of the χ_2^2 distribution.

Now let's assume that only X_1 is available to the observer. In this case

$$\begin{aligned}\eta(X_1) = P\{Y = 1|\ X_1\} =& P\{X_1^2 + X_2^2 + X_3^2 < c|\ X_1\}\\ =& P\{X_2^2 + X_3^2 < c - X_1^2|\ X_1\}\\ =& \max(0, F_2(c - X_1^2)),\end{aligned}$$

where F_2 denotes the cdf of χ_2^2 distribution. Thus Bayes decision function is

$$g^*(X_1) = \begin{cases} 1 & \text{if} \quad X_1^2 < c - F_2^{-1}(1/2) = 0.9797 \\ 0 & \text{otherwise.} \end{cases}$$

The probability of error is

$$\begin{aligned}L^* =& P\{g^*(X_1) \neq Y\} = P\{X_1^2 + X_2^2 + X_3^2 < c, X_1^2 > c - F_2^{-1}(1/2)\} +\\ & P\{X_1^2 + X_2^2 + X_3^2 \geq c, X_1^2 \leq c - F_2^{-1}(1/2)\}\\ =& \int_{c-F_2^{-1}(1/2)}^{c} F_2(c-x)f_1(x)dx + \int_0^{c-F_2^{-1}(1/2)} (1 - F_2(c-x)f_1(x)dx,\\ =& 0.3062\end{aligned}$$

where f_1 denotes pdf of the χ_1^2 distribution. Tables 8 and 9 provide simulational results, the confirm theoretical findings, as empirical error converges to theoretical Bayes error as sample size increases. Figure 7 gives graphical illustration of the results of the simulations.

Method	100	250	500	1000
WBGLC	0.2076	0.1876	0.1861	0.1828
$\text{WBGLC}_{\text{reg}}$	0.2158	0.2207	0.1805	0.1799
WBGLC_{MS}	0.2299	0.1944	0.1827	0.1756
Logistic Regression	0.2037	0.2046	0.2035	0.2033

Table 8: Example 2: X_1 and X_2 are observable. Average of the empirical errors of the 100 simulations. Wavelet function is Daubechies with 8 vanishing moments. The Bayes error is 0.1702

2.5.4 Effect of the Wavelet Function on the Performance of Classifier

Different wavelets have different properties. For example, Daubechies filters are minimal phase filters that generate wavelets which have minimal support for a given number of vanishing moments. Symmlets are within minimum size support for a given number of vanishing moments, but they are as symmetrical as possible, as apposed to the Daubechies filters which are highly asymmetrical. The

Method	100	250	500	1000
WBGLC	0.3792	0.3519	0.3220	0.3195
$\text{WBGLC}_{\text{reg}}$	0.4481	0.4348	0.3775	0.3367
WBGLC_{MS}	0.3722	0.3442	0.3227	0.3151
Logistic Regression	0.3414	0.3452	0.3404	0.3402

Table 9: Example 2: Only X_1 is observable. Average of the empirical errors of the 100 simulations. Wavelet function is Daubechies with 8 vanishing moments. The Bayes error is 0.3062

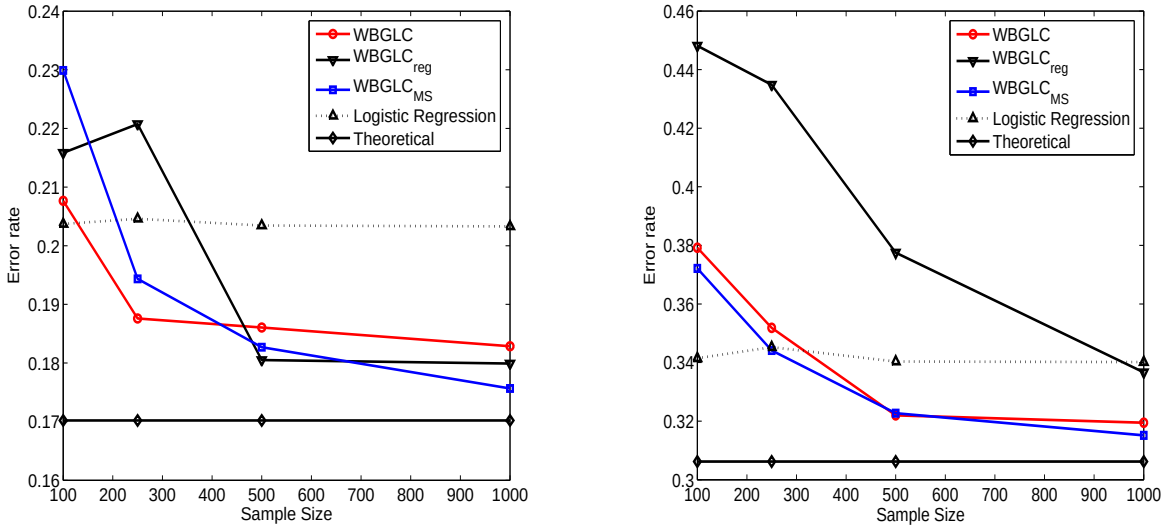


Figure 7: Example 2: Average of the empirical errors of the 100 simulations, for various training sample sizes. (a) Case 1: X_1 and X_2 are observable. (b) Case 2: Only X_1 is observable.

Vaidyanathan filter gives exact reconstruction, but does not satisfy any moment condition. Wavelets designed for a different purposes, for example, some are specifically designed for image processing, others for speech coding. So it is expected that that selection of the wavelet function will have impact on the performance of the proposed wavelet-based generalized classifier.

Our wavelet-based classifier approximates generally unknown function

$$\eta(x) = P\{Y = 1|X = x\} = \mathbb{E}\{Y|X = x\}.$$

Choice of wavelet function will determine how well we approximate $\eta(X)$. We now provide experimental results that confirm that selection of the wavelet function plays important role in performance of the proposed classifier.

Wavelet	100	200	300	500
Case 1: X_1 and X_2 are observable				
Daubechies 2	0.4268	0.4175	0.4221	0.3975
Daubechies 3	0.3480	0.3063	0.2962	0.2611
Daubechies 4	0.2827	0.2212	0.2143	0.1960
Daubechies 5	0.2377	0.1959	0.1870	0.1773
Daubechies 6	0.2367	0.1916	0.1832	0.1771
Daubechies 7	0.2419	0.1944	0.1875	0.1798
Daubechies 8	0.2448	0.1964	0.1889	0.1792
Daubechies 9	0.2572	0.1977	0.1900	0.1802
Daubechies 10	0.2608	0.2017	0.1937	0.1803
Case 2: Only X_1 is observable				
Daubechies 2	0.4261	0.3974	0.3927	0.3628
Daubechies 3	0.3870	0.3441	0.3236	0.3071
Daubechies 4	0.3671	0.3242	0.3082	0.2963
Daubechies 5	0.3609	0.3287	0.3100	0.3057
Daubechies 6	0.3610	0.3299	0.3193	0.3057
Daubechies 7	0.3620	0.3349	0.3201	0.3085
Daubechies 8	0.3677	0.3382	0.3211	0.3071
Daubechies 9	0.3756	0.3436	0.3244	0.3101
Daubechies 10	0.3756	0.3397	0.3259	0.3143

Table 10: Average of the empirical errors of the 50 simulations for the Daubechies family, for various training data sample sizes. The Bayes errors are 0.1710 and 0.2829, for case 1 and 2 respectively.

Tables 10 – 12, illustrate performance of the generalized linear classifier with multiples scales ($[5, 6, 7]$) on the data set described in the Example 1. For the same date we calculate empirical error of the classifier using different wavelet families, for for each family we also investigate effect of the number of vanishing moments. Results indicate that average of the empirical error over 50 simulation decreases as we increase number of vanishing moments and starts increasing after for number of vanishing moments greater that 6. Figure 8, demonstrates average of the empirical errors of the 50 simulations, for various training sample sizes and various wavelet function with 6 vanishing moments. Results indicate that Daubechies family outperforms Symmlet and Coiflet.

2.5.5 Application in Paper Producing Process

We consider an example from the book of Pandit and Wu ([62], pp. 496–497) which presents 100 data points of the observed basis weights in response to an input in the stock flow rate of a paper-making process. The values were taken at one-second intervals. The following brief description of

Wavelet	100	200	300	500
Case 1: X_1 and X_2 are observable				
`Coiflet 4`	0.3577	0.3208	0.3100	0.2817
`Coiflet 6`	0.2702	0.2221	0.2110	0.2107
`Coiflet 8`	0.2928	0.2472	0.2262	0.2267
`Coiflet 10`	0.3272	0.2704	0.2582	0.2565
Case 2: Only X_1 is observable				
`Coiflet 4`	0.4145	0.3809	0.3618	0.3543
`Coiflet 6`	0.4063	0.3963	0.3800	0.3714
`Coiflet 8`	0.4247	0.4122	0.3942	0.3870
`Coiflet 10`	0.4541	0.4276	0.4225	0.4088

Table 11: Average of the empirical errors of the 50 simulations for the Coiflet family, for various training data sample sizes. The Bayes errors are 0.1710 and 0.2829, for case 1 and 2 respectively.

Wavelet	100	200	300	500
Case 1: X_1 and X_2 are observable				
`Symmlet 4`	0.2870	0.2335	0.2269	0.2069
`Symmlet 6`	0.2484	0.2147	0.2044	0.1961
`Symmlet 10`	0.2966	0.2512	0.2408	0.2431
Case 2: Only X_1 is observable				
`Symmlet 4`	0.3852	0.3512	0.3419	0.3324
`Symmlet 6`	0.3857	0.3764	0.3687	0.3591
`Symmlet 10`	0.4292	0.4127	0.4138	0.4052

Table 12: Average of the empirical errors of the 50 simulations for the Symmlet family, for various training data sample sizes. The Bayes errors are 0.1710 and 0.2829, for case 1 and 2 respectively.

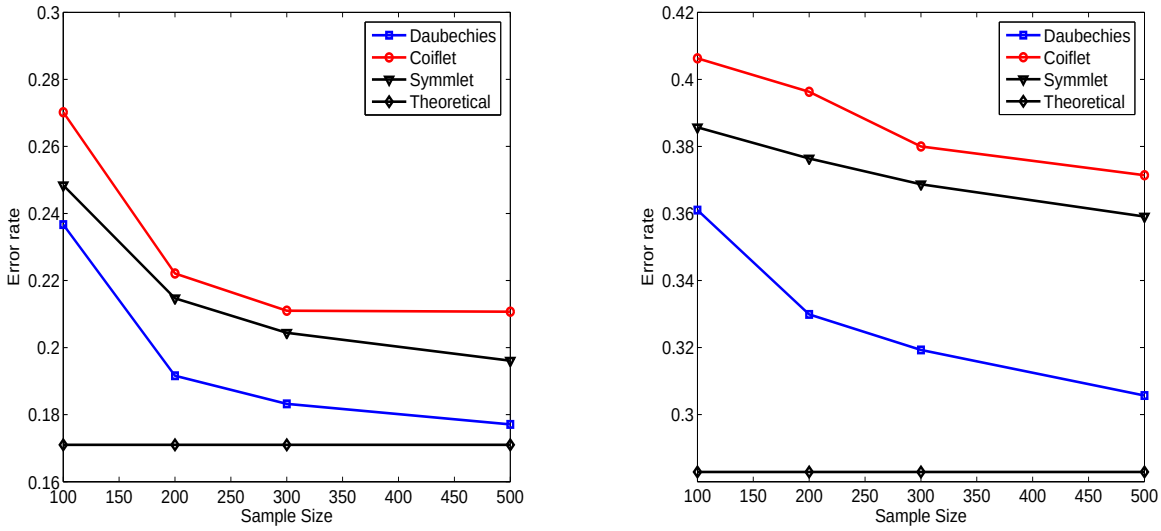


Figure 8: Average of the empirical errors of the 50 simulations, for various training sample sizes and various wavelet function with 6 vanishing moments. (a) Case 1: X_1 and X_2 are observable. (b) Case 2: Only X_1 is observable.

the papermaking process is from section 11.1.1 of Pandit and Wu [62]. A schematic diagram can be found there, too.

> The Fourdrinier papermaking process starts with a mixture of water and wood fibers (pulp) in the mixing box. The gate opening in the mixing box can be controlled to allow a greater or smaller flow of the thick stock (a mixture of water and fiber) entering the headbox. A turbulence is created in the headbox by means of suspended plates to improve the consistency of the pulp. The pulp then descends on a moving wire screen, as a jet from the headbox nozzles. Water is continuously drained from the wet sheet of paper so formed on the wire screen. The paper sheet then passes through press roles, driers, and calender roles to be finally wound.
>
> It is important to produce paper of as uniform a thickness as possible since irregularities on the surface such as ridges and valleys cause trouble in later operations such as winding, coating, printing, etc. This uniformity is measured by what is called a basis weight, the weight of dry paper per unit area. It may be measured directly or by means of a beta-ray gauge that makes use of the beta-ray absorption properties of the paper

material. The regulation of paper basis weight is one of the major goals of the paper control system.

The basis weight is affected by variables such as stock consistency, stock flow, headbox turbulence and slice opening, drier steam pressure, machine speed, etc. However, thick stock flow is often used as the main control input measured by the gate opening in the mixing box.

Based on the above description, we selected the stock flow as the only input in our problem. We are looking for a good predictor of the output basis weight. Let B_t and S_t for $t = 1, 2, \ldots, 100$ denote the basis weight and the stock flow rate at time t, respectively. The output basis weight, B_t depends not only on its past values but also on the stock flow. However, as stock flow must go through several steps such as refining, pressing, drying etc. *to be paper products,* the stock flow at time t cannot directly affect the basis weight at the same time. However, we assumed that S_{t-1} affects B_t. Further analysis found that $0.7B_{t-1} + 0.25S_{t-1}$ is a good predictor of B_t.

Now we define $\{(X_t, Y_t), t = 2, 3, \ldots, 100\}$. The target basis weight depends upon the grade of the paper being made. We assume that the target basis weight for the paper is $40\text{lb}/3300$ sq ft. and our tolerance level is ± 0.5lb. Therefore, we consider the basis weight, B_t in the range of 39.5 and 40.5 as "good" and assign the value of "1" for the response variable, Y_t. Otherwise, the basis weight is "bad" and Y_t is assigned "0". For each such Y_t, the corresponding X_t is $0.7B_{t-1} + 0.25S_{t-1}$. Thus, we have 99 data points of (X_t, Y_t)'s from the given 100 values of basis weight and stock flow. We used (X_t, Y_t)'s with odd t as the training set and the remaining even-index set as the validation set.

By identifying the classifier, we hope to be able to predict whether the future basis weight will be "good" or "bad" at the measured basis weight and stock flow. In addition, we want to make the output basis weight maintained at the "good" range of target value by manipulating the stock flow. For example, by looking at the measured basis weight and stock flow at time t, we can guess the basis weight at time $t + 1$ and from this future basis weight, we know which range of stock flow rate we should have to get a "good" basis weight at time $t + 2$.

The empirical errors for the wavelet classifier with different wavelet basis are given in Table 13.

Wavelet	WBGLC ($J = 4$)	WBGLC$_{\text{reg}}$ ($J = 4$)	WBGLC ($J = 4$)	WBGLC$_{\text{reg}}$ ($J = 5$)
`Symmlet 8`	0.36	0.34	0.38	0.26
`Daubechies 4`	0.24	0.20	0.28	0.20
`Daubechies 8`	0.22	0.16	0.30	0.18

Table 13: Empirical errors for the paper making process.

Chang, Kim, and Vidakovic [13] report an error of 0.18 for the same data set. The empirical error of the logistic regression is

$$L_{49}^{\text{logit}}(50) = \frac{1}{50}\sum_{j=1}^{50}\mathbf{I}\Big(\mathbf{I}(f(X_j) > 0.5) \neq Y_j\Big) = 0.36.$$

The best performance is achieved with Daubechies 8 wavelet function and regularized classifier, with error 0.16.

2.5.6 MicroArray Example

MicroArrays. From Wikipedia (`http://en.wikipedia.org/`), the free encyclopedia. A DNA microarray is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. The affixed DNA segments are known as reporters, thousands of which can be used in a single DNA microarray. Microarray technology evolved from Southern Blotting[1], where fragmented DNA is attached to a substrate[2] and then probed with a known gene or fragment. Measuring gene expression using microarrays is relevant to many areas of biology and medicine, such as studying treatments, disease and developmental stages.

The most common use of microarrays is to quantify mRNAs[3] transcribed[4] from different genes and which encode different proteins. RNA is extracted from many cells, ideally from a single cell type, then converted to cDNA or cRNA (complimentary DNA or RNA). Fluorescent tags [5] are enzymatically incorporated into the newly synthesized cDNA/cRNA or can be chemically attached to

[1]Southern Blotting is a method in molecular biology of enhancing the result of separation of the DNA strands by size, by marking specific DNA sequences.

[2]Substrate is a molecule which is acted upon by an enzyme.

[3]Messenger RNA (mRNA) is RNA that encodes and carries information from DNA to sites of protein synthesis.

[4]Transcription is the process through which a DNA sequence is copied by an RNA polymerase to produce a complementary RNA. Or, in other words, the transfer of genetic information from DNA into RNA.

[5]A fluorescent tag is a part of a molecule that researchers have attached chemically to aid in detection of the molecule to which it has been attached.

the new strands of DNA or RNA. A cDNA or cRNA molecule that contains a sequence complementary to one of the single-stranded probe sequences on the array will hybridize, via base pairing (more at DNA), to the spot at which the complementary reporters are affixed. The spot will then fluoresce (or glow) when examined using a microarray scanner.

Increased or decreased fluorescence intensity indicates that cells in the sample have recently transcribed, or ceased transcription, of a gene that contains the probed sequence (”recently,” because cells tend to degrade RNAs soon after transcription). The intensity of the fluorescence is roughly proportional to the number of copies of a particular mRNA that were present and thus roughly indicates the activity or expression level of that gene. Arrays can paint a picture or ”profile” of which genes in the genome are active in a particular cell type and under a particular condition.

The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. The large number of genes present on a single array means that the experimenter must take into account the multiple testing problem: even if each gene is extremely unlikely to randomly yield a result of interest, the combination of all the genes is likely to show at least one or a few occurrences of this result which are false positives.

microRNA. From Wikipedia (`http://en.wikipedia.org/`), the free encyclopedia. The term miRNA was first introduced in a set of three articles in Science (26 October 2001). In genetics, a miRNA (micro-RNA) is a form of single-stranded RNA which is typically 20-25 nucleotides long, and is thought to regulate the expression of other genes. miRNAs are RNA genes which are transcribed from DNA, but are not translated into protein. The DNA sequence that codes for an miRNA gene is longer than the miRNA. This DNA sequence includes the miRNA sequence and an approximate reverse complement. When this DNA sequence is transcribed into a single-stranded RNA molecule, the miRNA sequence and its reverse-complement base pair to form a double stranded RNA hairpin loop; this forms a primary miRNA structure (pri-miRNA).

The function of miRNAs appears to be in gene regulation. For that purpose, a miRNA is complementary to a part of one or more messenger RNAs (mRNAs), usually at a site in the 3’ UTR (prime untranslated region). The annealing of the miRNA to the mRNA inhibits protein translation. In some cases, the formation of the double-stranded RNA through the binding of the miRNA

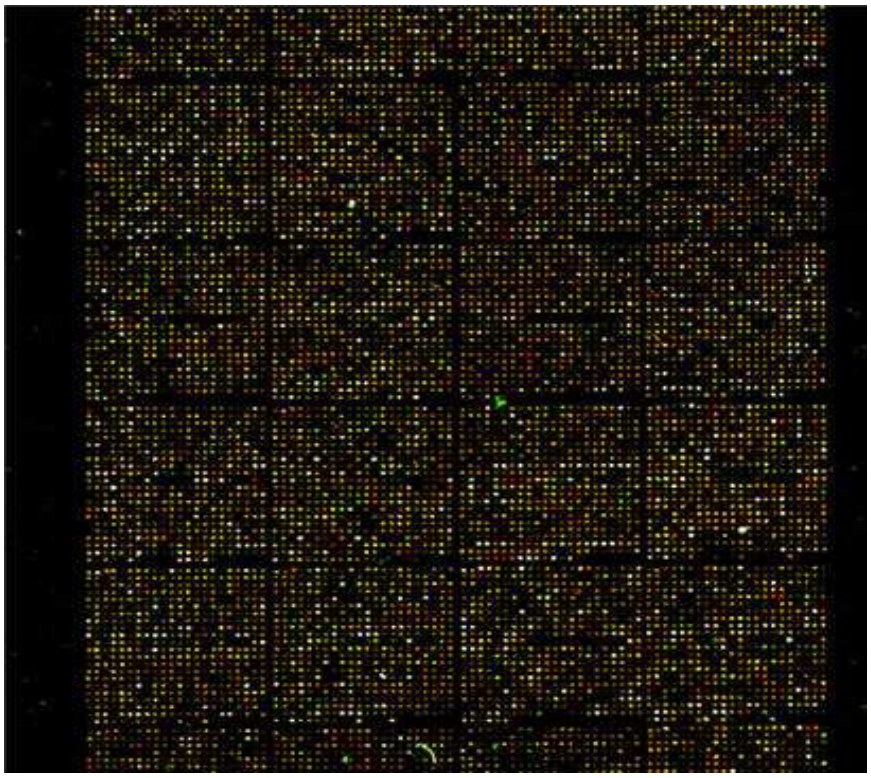


Figure 9: A DNA microarray, the different colors indicate relative expression of different genes. Image is taken from Wikipedia, the free encyclopedia.

triggers the degradation of the mRNA transcript through a process similar to RNA interference (RNAi), though in other cases it is believed that the miRNA complex blocks the protein translation machinery or otherwise prevents protein translation without causing the mRNA to be degraded.

miRNA and cancer.

miRNA has been found to have links with some types of cancer. A study of mice altered to produce excess c-myc a protein implicated in several cancers shows that miRNA has an effect on the development of cancer. Mice that were engineered to produce a surplus of types of miRNA found in lymphoma cells developed the disease within 50 days and died two weeks later. In contrast, mice without the surplus miRNA lived over 100 days. (He, et al., 2005)

Another study found that two types of miRNA inhibit the E2F1 protein, which regulates cell proliferation. miRNA appears to bind to messenger RNA before it can be translated to proteins that switch genes on and off. (O'Donnell, et al., 2005)

By measuring activity among 217 genes encoding miRNA, patterns of gene activity that can

distinguish types of cancers can be discerned. miRNA signatures may enable classification of cancer. This will allow doctors to determine the original tissue type which spawned a cancer and to be able to target a treatment course based on the original tissue type. miRNA profiling has already been able to determine whether patients with chronic lymphocytic leukemia had slow growing or aggressive forms of the cancer. (Lu, et al., 2005)

References:

- This paper defines miRNA and proposes guidelines to follow in classifying RNA genes as miRNA: Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T (2003). A uniform system for microRNA annotation. RNA 9 (3): 277-279. PMID 12592000

- This paper discusses the processes that miRNA and siRNAs are involved in, in the context of 2 articles in the same issue of the journal Science: Baulcombe D (2002). DNA events. An RNA microcosm.. Science 297 (5589): 2002-2003. PMID 12242426

- This paper describes the discovery of lin-4, the first miRNA to be discovered (editor's note: in fact, no Wikipedia editor has yet read this paper, only made inferences from a citation): Lee RC, Feinbaum RL, Ambros V (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75 (5): 843-854. PMID 8252621

- He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM (2005). A microRNA polycistron as a potential human oncogene. Nature 435 (7043): 828-833. PMID 15944707

- O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature 435 (7043): 839-843. PMID 15944709

- Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR (2005). MicroRNA expression profiles classify human cancers. Nature 435 (7043): 834-838. PMID 15944708

Data Description. In their paper, Lu et al. [43], used miRNA for classification of cancer. They used miRNA expressions in the human samples as a training sample and successfully classified mouse samples using same miRNA expressions. Paper, supplementary materials and data sets can be found at Cancer Program Data Sets at Broad Institute at

`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`.

In order to build classifier of normal samples vs. tumor samples based on miGCM collection, we selected tissues used by Lu et al. [43]. Table 14 summarizes the tissues for the analysis. Testing data set mLung consisted of the 12 mouse lung samples (5 normal and 7 with tumor). Both data sets have 217 miRNA expressions. Each data set was $\log_2$ transformed.

Tissue	Number of Normal	Number of Tumor
Colon	5	10
Kidney	3	5
Prostate	8	6
Uterus	9	10
Lung	4	6
Breast	3	6

Table 14: Number of training samples used to build Normal/Tumor classifier.

Modeling. Proposed above wavelet-based generalized linear classifier learns from training data set (above samples from miGCM data set) and predicts samples in a test data set (the mouse lung sample set). Wavelet-based classifier with `Daubechies 6` wavelet function and scale set $[3, 4, 5]$, was applied individually to miRNA expression. Each expression has 75 observation, each observation can be classified into class 0 (normal sample) and class 1 (tumor sample). A set of 132 markers (miRNA expressions with small error rate (less that 30%), i.e. that best distinguishes between two classes of samples) was selected using training data set. 109 of those selected markers is identical to the ones selected by Lu et al. [43]. These 132 markers were used without modification to predict 12 mouse lung samples. Each mouse sample was predicted separately, using wavelet-based generalized linear classifier. Since, as mentioned above, each individual expression is unlikely to yield

results of interest, classification is done according to the following rule:

$$\text{Class} = \begin{cases} 0 & \text{If } \frac{\sum_{i=1}^{132} g_{75}(X_i, D_{75}^i)}{132} < 0.5 \\ 1 & \text{otherwise.} \end{cases}$$

where X_i is miRNA expression for marker i, D_{75}^i learning set of 75 human samples corresponding to i^{th} miRNA expression. Using this procedure 11 out of 12 samples were correctly identified , Lu et al. [43] reported 100% accuracy, but they used *k*-NN (*k*-nearest neighbor classifier) which is much more powerful method.

2.6 Conclusion

In this chapter we introduce wavelet-based generalized linear classifier. We establish that under mild conditions this classifier is both consistent and strongly consistent. Experimental results show that the proposed classifier performs well and is comparable to other methods.

MATLAB programs for the wavelet-based generalized linear classifier are available at Jacket's Wavelets page `http://www.isye.gatech.edu/~brani/wavelet.html`.

CHAPTER III

LOCAL BAYESIAN FALSE DISCOVERY RATE WAVELET SHRINKAGE

3.1 Chapter Introduction

In this chapter we introduce wavelet-based shrinkage based on two versions of false discovery rate: local FDR and Bayesian FDR based on selecting dominant posterior probabilities. The developed methodology is comparable to currently best available wavelet shrinkage methods. Even though the two proposed methods may not achieve the minimum of MSE they possess two distinct qualities: (i) they are of thresholding type leading to most parsimonious representations desirable when dimension reduction is an issue, and (ii) the bias of obtained estimators is small.

Simultaneous testing of multiple hypotheses has always attracted the attention of statisticians (e.g., Folks [29]) but traditionally, the number of hypotheses was modest (say, < 20). Nowadays, thousands of hypotheses need to be tested simultaneously and the traditional methods (such as Bonferroni, for example) are not sensible because of loss of specificity and power.

To illustrate the loss of specificity, consider a gene expression example. Assume that a chip contains 10000 genes and not a single gene is differentially expressed. If we test each of 10000 hypotheses separately at level $\alpha = 0.01$, we would expect that 10000 $\times 0.01 = 100$ of the tests would have p-value less than α, i.e., 100 of the tests would be falsely significant and the probability that at least one p-value will be less than α (family-wise error rate) is around 1. Thus, individual p-values are no longer valid measures of significant findings.

For controlling the FWER (family-wise error rate, Dudoit *et al.* [25]), conservative methods such as Bonferroni correction is widely used, however this method also suffers from the lack of power when the number of hypotheses is large. For microarray data, for example, the goal is to focus on several candidate genes for further study. Thus, the low power of FWER-controlling procedure is unacceptable and it would be better to control the false discovery rate, FDR, a method

that is discussed in some detail in following Sections.

To formally illustrate what happens in a testing problem when the number of hypotheses to be tested simultaneously increases, we consider paradigmatic problem of testing for the multivariate normal mean.

Suppose we wish to test

$$H_0 : \theta = 0 \ \ vs. \ \ H_1 : \theta \neq 0, \tag{33}$$

where $\mathbf{X} \ \sim \ \mathcal{MVN}_n(\theta, I_n)$ is observed. A sufficient statistics for the problem is $||\mathbf{X}||^2$. If the alternative is precise, $\theta = \theta_1$, then the α-level maximum likelihood ratio test has approximate power

$$1 - \Phi\left(\frac{z_{1-\alpha} - ||\theta_1||^2/\sqrt{2n}}{\sqrt{1 + 2||\theta_1||^2/n}}\right) \approx 1 - \Phi\left(z_{1-\alpha} - ||\theta_1||^2/\sqrt{2n}\right). \tag{34}$$

If $||\theta_1||^2$ goes to infinity, the power of the test is expected to tend to 1, however, if $||\theta_1||^2$ goes to infinity as $o(\sqrt{n})$, the power tends to α when n increases. Thus, for high-dimensional θ, the discriminatory distance $||\theta_1||^2$ is shrunk to $||\theta_1||^2/\sqrt{2n}$, and the power tends to the significance level.

This dissipation of power, while testing multiple hypotheses is discussed by many researchers. Folks [29] gives an excellent overview of multiple hypothesis testing, including the Tippet method which can be viewed as a precursor of FDR method of Benjamini and Hochberg [11], and local FDR discussed in Section 3.2. Another classical repository of methods used in multiple hypothesis testing is monograph by Miller [50].

In his spirited paper with applications in genomics, Efron [26] found that local FDR tends to overfit the model. He demonstrated that replacing the "theoretical null distribution" by its empirical counterpart often improves the model selection. We connect local FDR approach (Efron and Tibshhirani [27]; Efron [26]) to the related model selection procedure based on Bayes factors in the context of wavelet-smoothing. Consistently with Efron's findings, the empirical H_0 density in the local FDR tends to have longer tails, leading to more parsimonious models. An efficient proposal to replace theoretical null by empirical null based on nonparametric version of Empirical Bayes estimator is proposed by Datta and Datta [16].

3.2 Local False Discovery Rate in the Wavelet Domain (BLFDR)

Many proposed wavelet shrinkage methods can be interpreted as multiple hypotheses testing in the wavelet domain. For example the *universal thresholding* of Donoho and Johnstone [22], *recursive likelihood ratio tests* of Ogden and Parzen [61], *false discovery rate* of Abramovich and Benjamini ([2], [3]) are some early references. Vidakovic [72] proposes the use of Bayesian hypothesis testing and Bayes factors in the tasks of wavelet thresholding. Vidakovic and Ruggeri [74] develop an adaptive Bayesian model in which the resulting Bayes rule acts as a shrinker in the wavelet domain. Their method (Bayesian Adaptive Multiscale Shrinkage, or short BAMS) is now part of *Gaussian-WaveDen* of Antoniadis, Bigot,and Sapatinas [6] and allows for incorporation of prior information about the signal. We review the local false discovery rate and establish the link with Bayes factor shrinkage induced by BAMS model, all in the context of wavelet shrinkage.

Suppose the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ represent the sum of an unknown signal $\boldsymbol{f} = (f_1, \ldots, f_n)$ and random noise $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. Coordinate-wise, $y_i = f_i + \epsilon_i,\ i = 1, \ldots, n$. In the wavelet domain (after applying a linear and orthogonal wavelet transformation $\boldsymbol{W}$ to the observed data), expression (35) becomes $d_{jk} = \theta_{jk} + \epsilon_{jk},\ i = 1, \ldots, n$, where d_{jk}, θ_{jk}, and ϵ_{jk} are the j, k-th coordinates in the traditional scale/shift wavelet-enumeration of vectors $\boldsymbol{Wy}, \boldsymbol{Wf}$ and $\boldsymbol{W\epsilon}$, respectively. Our assumption is that the coefficients d_{jk} can be considered independently, since the wavelet transformations are decorrelating. When modeling in practice, such an assumption prove to be very reasonable. In the exposition that follows, we omit the double index jk and work with a "typical" wavelet coefficient, d. Therefore, our model is

$$d = \theta + \epsilon. \tag{35}$$

One way to select the parsimonious model is to componentwise test that the signal part of the coefficient is zero, i.e., $H_0 : \theta = 0$. If the hypothesis is rejected the coefficient is significant and retained in the model. If the H_0 is accepted, then d in the model is replaced by 0. After all n tests are conducted, the coefficients that survived the tests are back transformed to the domain of the original data.

When n simultaneous null hypotheses are tested, the corresponding test statistics (likely not all independent) will result in n p-values, $p_1, p_2, \ldots, p_n$. Under H_0 these p-values represent a sample

from a uniform distribution.

It is more convenient to work with z-values,$z_1, \ldots, z_n$, where

$$z_i = \Phi^{-1}(p_i). \tag{36}$$

Under the H_0 the z_is are theoretically $N(0,1)$, and the standard normal tables could be used.

Define local false discovery rate as the ratio of $f_0(z)$, theoretical null density for zs and $f(z)$ observed empirical density for z's,

$$\textbf{lfdr}(z) = \frac{f_0(z)}{f(z)}.$$

Efron [26] suggests to keep in the model, as *interesting*, all coefficients for which $\textbf{lfdr}(z_i) = \frac{f_0(z_i)}{f(z_i)}$ is smaller than some threshold value, say $\gamma = 0.10$. As pointed in the same paper, by dropping p_0 which is close to 1 (most of the H_0s are true, only a few coefficients are retained in the model), the probability $P(\text{Uninteresting}|z) = p_0 f_0(z)/f(z)$ is close to $\textbf{lfdr}(z) = f_0(z)/f(z)$ and represents a link with Bayes factor shrinkage proposed by Vidakovic [72].

Next we introduce the local false discovery rate for a specific model first discussed in the wavelet context by Vidakovic and Ruggeri [74].

Assume that $[d|\theta, \sigma^2] \sim N(\theta, \sigma^2)$ and the prior $\sigma^2 \sim \mathcal{E}(\mu), \ \ \mu > 0$, with density $f(\sigma^2|\mu) = \mu e^{-\mu\sigma^2}$. The marginal likelihood (with σ^2 integrated out) is

$$[d|\theta] \sim \mathcal{DE}\left(\theta, \frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta) = \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}.$$

If the prior on θ is

$$[\theta] \sim \pi_o\delta_0 + \pi_1\mathcal{DE}(0,\tau), \tag{37}$$

then the predictive distribution of d is

$$[d] \sim m(d) = \pi_0\mathcal{DE}\left(0, \frac{1}{\sqrt{2\mu}}\right) + \pi_1 m_1(d),$$

where $m_1(d)$ is

$$\frac{\tau e^{-|d|/\tau} - \frac{1}{\sqrt{2\mu}}e^{-\sqrt{2\mu}|d|}}{2\tau^2 - 1/\mu},$$

The Bayes factor in favor of testing $H_0 : \theta = 0$ versus the alternative $H_1 : \theta \neq 0$, when wavelet coefficient d is observed is

$$B_{01}(d) = \frac{f(d|0)}{m_1(d)} = \frac{\frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d|}}{\frac{\tau e^{-|d|/\tau} - \frac{1}{\sqrt{2\mu}}e^{-\sqrt{2\mu}|d|}}{2\tau^2 - 1/\mu}} \tag{38}$$

By straightforward calculation,

$$f_0/m(d) = \frac{B_{01}(d)}{\pi_1 + \pi_0 B_{01}(d)} \tag{39}$$

and **lfdr** can be estimated by $\hat{B}_{01}(d)/(\pi_1 + \pi_0\hat{B}_{01}(d))$ where $\hat{B}_{01}$ is an empirical counterpart of Bayes factor.

Using the empirical counterpart of Bayes factor means that we choose the empirical null hypothesis instead of theoretical one. Local false discovery rate, lfdr, computed by using the empirical counterpart of Bayes factor ($\hat{B}_{01}$) corresponds to the empirical null hypothesis and lfdr computed by using the exact Bayes factor (B_{01}) corresponds to the theoretical null hypothesis.

In multiple testing problem, Efron [26] considered the choice of an appropriate density for the null hypotheses, the point there being that large-scale situations can provide their own "empirical null", which may differ in important ways from the traditional theoretical null appropriate for any individual problem. In addition, permutation and bootstrap null density estimates should be considered as improved versions of theoretical null, rather than empirical nulls.

Remark 1. In the illustrative example of gene expression data (Jung et al., [40]), it is shown that the frequently used permutation methods can be misleading when the mean of the distribution of test statistics for most genes(non-differential genes) is not zero because the random permutations of expression levels across the control and treatment groups make the mean of the distribution definitely zero. This would yield a bias in the mean estimate and thus result in inaccurate estimation of FDR.

Jung et al. [40] proposed the fully Bayesian mixture model-based method in meta-analysis to estimate the null distribution of test statistics and compared it with the permutation methods by computing the FDRs given the critical value. The proposed method was applied to four publicly available prostate cancer gene expression data and the results showed that the model-based approach is superior to the permutation method.

In this example, the mean of test statistics of non-differential genes was estimated 0.177, larger than zero and the null density from permutation method showed the center around zero, showing significant difference from the estimated density in the proposed method. Also for example, with critical value 1.207, the FDR estimated by the mixture model is 0.001, while the FDR estimated by the permutation method is 0.022.

Remark 2. From the form of Bayes factor in (38), it follows that $B_{01} < \alpha$ leads to a thresholding rule $|d| > \lambda$, for some $\lambda = \lambda(\alpha)$.

The Bayes factor provides a measure of data support for the H_0 and is used to calculate posterior odds of H_0 as

$$\frac{p_0}{p_1} = \frac{\pi_0}{\pi_1} B_{01},$$

where $\frac{\pi_0}{\pi_1}$ are prior odds.

A coefficient should remain in the model if the $B_{01} < \alpha$. Since adaptive Bayesian shrinkage uses level varying probability of null hypothesis, $\pi_0 = \pi_0(j)$, where j is the level in the wavelet decomposition, the local false discovery rate is equivalent to the following rule based on Bayes factors,

Keep the wavelet coefficient at level j in the model as interesting if $B_{01} \leq \alpha(j)$.

In automatic procedure $\alpha(j)$ is always 1, which reflects the fact that H_0 is more readily rejected. This means that the thresholding is not performed and the coefficient is retained as significant.

3.3 FDR Ordering of Posterior Probabilities (BaFDR)

As we hinted in the Introduction, wavelet shrinkage, in form of thresholding of wavelet coefficients, can be viewed as a multiple testing problem. For each observed wavelet coefficient $d_i = \theta_i + \sigma\epsilon_i$, consisting of signal part θ_i and the error $\sigma\epsilon_i$ the hypothesis $H_0 : \theta_i = 0$ is tested against the alternative $H_1 : \theta_i \neq 0$. If the hypothesis H_0 is rejected, the coefficient d_i is retained in the model as significant. Otherwise, it is discarded.

For example, the universal threshold can be viewed as a critical value of a test with the level

$$\alpha = P(|d_i| > \sqrt{2\log n}\,\sigma \,|H_0\,) \approx (n\sqrt{\pi \log n})^{-1}.$$

The power of this test against the alternative $H_1 : \theta_i = \theta\ (\neq 0)$ is $O\left(\frac{1}{n\sqrt{\log n}}\right)$.

Universal thresholding is equivalent to a Bonferroni-type procedure. In testing n statistical hypotheses simultaneously, the Bonferroni procedure guarantees that the overall level of the omnibus test is α by setting the levels for the individual hypotheses as $\frac{\alpha}{n}$. For large n, the individual levels $\frac{\alpha}{n}$ become unduly small, leading to loss of "strictness" and dissipation of power. This loss of strictness means that many of $H_0 : \theta = 0$ are accepted, i.e., many observed coefficients are discarded from the model leading to over-smoothing.

A way to control such dissipation of massive acceptance of null hypotheses could be based on the *false discovery rate* (FDR) (Abramovich and Benjamini [2], [3]; Benjamini and Hochberg [11]).

Here is a brief description. Let R be the number of wavelet coefficients retained in the model. If S of them are correctly kept, then $V = R - S$ are erroneously kept. The random variable $Q = V/R$ expresses the error in such a procedure. The false discovery rate of coefficients is the expectation of Q; that is, the expected proportion of coefficients erroneously kept. One maximizes the number of coefficients kept, subject to condition $\mathbb{E}Q \leq \alpha$, for α small.

Several Bayesian alternatives to FDR are proposed from the Bayesian stand point, a nice overview can be found in Tadesse *et al.* [69].

Rosner and Vidakovic [67] propose an FDR procedure in which that is based on the assessment of posterior probabilities of hypotheses. An application is given in Angelini and Vidakovic [5]. Suppose that in testi([18], [19])ng of n hypotheses H_0, we obtain a sequence of their posterior probabilities, $p_0^1, p_0^2, \ldots, p_0^n$. Let $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ be increasingly ordered posterior probabilities, and $q_{(k)} = 1 - p_{(k)},\ k = 1, \ldots, n$.

When deciding about retaining the wavelet coefficients in the model ("a discovery", "interesting coefficient", etc.) by rejecting corresponding null hypotheses $H_0 : \theta = 0$, one controls the number of hypotheses that are erroneously rejected, V. If the R hypotheses with smallest posterior probabilities are rejected, we require the expectation (with respect to the posterior measure) of $Q = V/R$ not to exceed α. Note that

$$\begin{aligned} \mathbb{E}Q &= \frac{1}{R}\sum_{i=0}^{R} iP(\text{Among } R \text{ rejected hypotheses, the number of erroneously rejected is } i) \\ &= \frac{1}{R}\sum_{i=0}^{R} iP_R(i), \end{aligned} \tag{40}$$

where the probabilities $P_R(i)$ can be calculated efficiently as the coefficients with powers z^i in generating polynomial

$$\varphi_R(z) = \prod_{k=1}^{R}(q_{(k)} + p_{(k)}z) = \sum_{i=0}^{R} P_R(i)z^i. \tag{41}$$

Thus, this Bayesian FDR procedure (BaFDR) can be summarized as follows:

• **STEP 1.** Find the posterior probabilities p_{jk} of all hypotheses $H_0 : \theta_{jk} = 0$ and order them according to their size.

• **STEP 2.** Fix α small and set $R = 1$.

• **STEP 3.** Increase R by 1. Find $\varphi_R(z)$ using $p_{(1)}, \ldots, p_{(R)}$, and calculate $\mathbb{E}Q$.

• **STEP 4.** If $\mathbb{E}Q \geq \alpha$ then the maximum posterior probability of rejection is $p_{(R-1)}$. **STOP.**

Otherwise, if $\mathbb{E}Q < \alpha$, return to **STEP 3.**

The introduced BaFDR naturally leads to wavelet thresholding. It turns out that such shrinkage is also linked with the shrinkage based on Bayes Factors and **lfdr** discussed in the previous sections.

Note that the posterior probability p_0 of hypothesis H_0 is

$$p_0(d) = \frac{B_{01}(d)}{\frac{\pi_1}{\pi_0} + B_{01}(d)} \tag{42}$$

where d is observed wavelet coefficient and $\frac{\pi_1}{\pi_0}$ are prior odds in favor of H_1. This is an easy reformulation of the definition of Bayes Factor which links the prior and posterior odds:

$$\frac{p_0}{p_1} = B_{01} \times \frac{\pi_0}{\pi_1}.$$

If the hypotheses H_0 is rejected, by (42),

$$p_0(d) \leq \alpha \quad \text{is equivalent to} \quad B_{01}(d) \leq \frac{\alpha}{1-\alpha} \times \frac{\pi_1}{\pi_0}.$$

We provide the simulational results involving the standard test functions and the BaFDR shrinkage. Because of its global nature the resulting shrinkage is inferior to the state of art local, neighborhood-dependent shrinkage methods.

3.4 Simulations and Application

The same setup is used for both BaFDR, and Local Bayesian FDR in wavelet domain. Four standard test functions (`blocks, bumps, doppler` and `heavisine`) are rescaled so that an added

standard normal noise produces a preassigned signal-to-noise ratio (SNR). The wavelet bases used are: Symmlet 8 for `doppler` and `heavisine`, Haar for `blocks` and Daubechies 6 for `bumps`, as standardly done. Number of levels in wavelet decomposition is 4 for signal length of 512, 5 levels for signal length 1024 and 6 for signal length 2048. The in all three cases the smooth level contains 32 coefficients which are left intact.

One of the key challenges in shrinkage/thresholding methods based on Bayesian model is specification of hyperparameters. It is desirable to have an automatic and objective procedure amenable to a range of input signals and noises. Our method is based on Empirical Bayes moment-matching. In principle it is possible to use more formal Empirical Bayes MLII method, but for practicable models such avenue leads to a nontrivial extremal problems.

We discuss two cases in specifying the hyperparameters. In the Case 1 the parameters are specified in a global way, i.e., coefficients in all detail levels have the same model. This case is compared to two popular global methods: VisuShrink and SureShrink (Donoho and Johnstone [22]; Donoho [23]; and Johnstone and Donoho, [39]).

In the Case 2 the model parameters depend on detail level, thus the models are level-dependent. The level dependent shrinkage is compared to ABWS of Chipman, Kolaczyk, and McCulloch [14] and BAMS of Vidakovic and Ruggeri [74]. Both of these methods are implemented by Antoniadis, Bigot, and Sapatinas [6]. More detailed description of this automatic hyperparameter selection is provided next.

3.4.1 Tuning the Model Parameters: Case 1

This is global model, i.e., hyperparameters in models for all detail coefficients are the same.

1. μ is the reciprocal of the mean for the prior on σ^2, or, equivalently, the square root of the precision for σ^2. We first estimate σ by a robust Tukey's $\texttt{pseudos} = (Q_1 - Q_3)/C$, where Q_1 and Q_3 are the first and the third quartiles of the finest level of details in the decomposition and $1.3 \leq C \leq 1.5$. We propose $\frac{1}{\texttt{pseudos}^2}$ as a default value for μ; according to the Law of Large Numbers, this ratio should be close to the "true" μ.

2. π_0 is the weight of the point mass at zero in the prior on θ and taken to be independent of level j.

3. τ is the scale of the "spread part" in the prior (37). In the case of a double exponential prior, the variance of the signal part is $2\tau^2$. Because of the independence between the error and the signal parts, we have $\sigma_d^2 = 2(1-\pi_0)^2\tau^2 + 1/\mu$, where σ_d^2 is the variance of the observations d. This yields

$$\tau = \sqrt{\max\left\{\frac{\sigma_d^2 - \frac{1}{\mu}}{2(1-\pi_0)^2}, 0\right\}}.$$

Note when $\tau = 0$, the prior (also the posterior) put all their mass at 0, which results in $\delta(d) = 0$.

3.4.2 Tuning the Model Parameters: Case 2

Models are level-dependent, i.e., some hyperparameters in models for detail coefficients are the same within a level, and different for different levels.

1. μ is specified as in the Case 1.

2. π_0 is the weight of the point mass at zero in the prior on θ and should depend on level j. Depending on our prior information about smoothness, π_0 should be close to 1 at the finest level of detail and close to 0 at the coarsest levels. We propose a hyperbolic decay in j,

$$\pi_0(j) = 1 - \frac{1}{(j - \texttt{coarsest} + 1)^\gamma}, \ \ \texttt{coarsest} \le j \le \log_2 n,$$

where `coarsest` is the coarsest level subjected to shrinkage.

3. Specification of τ coincides with that in Case 1 but with π_0 replaced by $\pi_0(j)$. In this case, $\sigma_d^2 = 2(1-\pi_0(j))^2\tau^2 + 1/\mu$, and

$$\tau = \sqrt{\max\left\{\frac{\sigma_d^2 - \frac{1}{\mu}}{2(1-\pi_0(j))^2}, 0\right\}}.$$

3.4.3 Results

Table 15 gives the mean-squared error MSE (Variance+Bias^2) for VisuShrink, SureShrink, BaFDR ($\alpha = 0.05$), and BLFDR-fixed, as procedures with a global shrinkage model and for BAMS, ABWS, and BLFDR-ld as level dependent shrinkers on standard test signals. The test signals are rescaled so that the noise variance σ^2 equals 1. Signal-to-noise ratio is 7 and sample size is 1024.

	blocks	bumps
VISUSHRINK	0.6840 (0.0719 + 0.6122)	1.5707 (0.1165 + 1.4543)
SURESHRINK	0.2225 (0.1369 + 0.0856)	0.6827 (0.2660 + 0.4167)
BAFDR	0.1460 (0.1137 + 0.0322)	0.5768 (0.2880 + 0.2888)
BLFDR-FIXED	0.1244 (0.1129 + 0.0115)	0.3796 (0.2584 + 0.1212)
ABWS	0.0995 (0.0874 + 0.0121)	0.3495 (0.2228 + 0.1267)
BAMS	0.1107 (0.0965 + 0.0142)	0.3404 (0.1976 + 0.1428)
BLFDR-LD	0.1184 (0.1154 + 0.0031)	0.3828 (0.2637 + 0.1191)
	doppler	heavisine
VISUSHRINK	0.4850 (0.0523 + 0.4327)	0.1204 (0.0339 + 0.0864)
SURESHRINK	0.2285 (0.0946 + 0.1340)	0.0949 (0.0416 + 0.0534)
BAFDR	0.2489 (0.1049 + 0.1440	0.1098 (0.0463 + 0.0635)
BLFDR-FIXED	0.1817 (0.1272 + 0.0545)	0.1010 (0.0689 + 0.0320)
ABWS	0.1646 (0.1006 + 0.0640)	0.0874 (0.0442 + 0.0433)
BAMS	0.1482 (0.0899 + 0.0584)	0.0815 (0.0511 + 0.0304)
BLFDR-LD	0.1801 (0.1283 + 0.0519)	0.1070 (0.0814 + 0.0256)

Table 15: MSE (Variance+Bias^2) for VisuShrink, SureShrink, BaFDR ($\alpha = 0.05$) and BLFDR (as global methods) and ABWS, BAMS, BLFDR (as level-wise methods). The standard test signals are rescaled so that the noise variance σ^2 equals 1. SNR is 7, and sample size is 1024.

Table 15 gives MSE (Variance+Bias^2) for VisuShrink, SureShrink, BaFDR ($\alpha = 0.05$) and BLFDR as comparable global methods. In addition to superior MSE, Bayesian hard-thresholding alternatives have much smaller bias.

To illustrate performance of BLFDR and BaFDR for standard signals and SNR's we provide three tables with simulational results. Tables 16 and 17 give global and levelwise BLFDR. For the global case $p_0 = 0.95$ while in the levelwise case parameters are determined as in the Case 2 with $\gamma = 2.5$. Table 18 gives MSE value for global shrinkage induced by BaFDR with $\alpha = 0.05$ and $\pi_0 = 0.90$.

Figures 10 – 13 show a graphical examples of the application of the above concepts.

Figure 14 shows ordered posterior probabilities (from BaFDR). Note that, as expected, for most of the coefficients the posterior probability is close to 1. On the other hand, the selection principle is robust with respect to the choice of maximal posterior probability – the number of coefficients in the model is essentially the same for all value of the posterior probability smaller than 0.9.

Function	n	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	512	0.2434	0.2159	0.1982	0.1810
	1024	0.1884	0.1433	0.1244	0.1044
	2048	0.1279	0.0904	0.0698	0.0570
Bumps	512	0.5022	0.5733	0.6596	0.7407
	1024	0.3356	0.3660	0.3796	0.3946
	2048	0.2235	0.2227	0.2261	0.2384
Doppler	512	0.2439	0.2524	0.2676	0.2872
	1024	0.1684	0.1692	0.1817	0.1901
	2048	0.1180	0.1044	0.1053	0.1079
Heavisine	512	0.1510	0.1593	0.1888	0.2123
	1024	0.1120	0.0943	0.1010	0.1185
	2048	0.0897	0.0688	0.0698	0.0796

Table 16: Performance of Local False Discovery Rate in Wavelet Domain. The table shows average MSE for 1000 simulations, with parameters τ and $\pi_0 = 0.95$ fixed for all levels.

3.4.4 An Application in AFM

To illustrate features of the BLFDR and BaFDR shrinkage approaches proposed here we used measurements in atomic force microscopy (AFM).

The AFM is a type of scanned proximity probe microscopy (SPM) that can measure the adhesion strength between two materials at the nanonewton scale (Binnig, Quate and Gerber, [12]). In AFM, a cantilever beam is adjusted until it bonds with the surface of a sample, and then the force required to separate the beam and sample is measured from the beam deflection. Beam vibration can be caused by factors such as thermal energy of the surrounding air or the footsteps of someone outside the laboratory. The vibration of a beam acts as noise on the deflection signal; in order for the data to be useful this noise must be removed.

The AFM data from the adhesion measurements between carbohydrate and the cell adhesion molecule (CAM) E-Selectin was collected by Bryan Marshal from the BME Department at Georgia Institute of Technology. The technical description is provided in Marshall, McEver, and Zhu [49].

Figure 15 depicts the original AFM signal (Panel (a)), signal smoothed by BaFDR procedure (Panel (b)), signal smoothed with global BLFDR procedure with $\pi_0 = 0.999$ fixed for all levels (Panel (c)), and signal smoothed by BLFDR with level-dependent π_0 but γ fixed at 5.

Function	n	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	512	0.2859	0.2483	0.2073	0.1688
	1024	0.1997	0.1472	0.1184	0.0977
	2048	0.1101	0.0910	0.0750	0.0595
Bumps	512	0.4876	0.5540	0.6174	0.6702
	1024	0.3272	0.3698	0.3828	0.3877
	2048	0.1981	0.2210	0.2381	0.2594
Doppler	512	0.2759	0.2916	0.2982	0.3049
	1024	0.1625	0.1699	0.1801	0.1912
	2048	0.0858	0.0942	0.1081	0.1178
Heavisine	512	0.1981	0.1834	0.1900	0.1966
	1024	0.1077	0.1010	0.1070	0.1202
	2048	0.0598	0.0566	0.0606	0.0700

Table 17: Performance of Local False Discovery Rate in Wavelet Domain. The table shows average MSE for 1000 simulations, with level-dependent parameters τ and π_0, $\gamma = 2.5$.

3.5 Conclusion

In this chapter we proposed and explored two natural approaches to threshold wavelet coefficients. The approaches are based on multiple testing of hypotheses in Bayesian fashion. They are linked with the hard thresholding paradigm and also with local false discovery rate methodology proposed and explored by Efron and Tibshirani [27] and Efron [26]. The proposed approaches are desirable when dimension reduction is important and they have small bias, as typical for hard-thresholding estimators.

The methodology leading to BLFDR is quite general and could be developed for a range of Bayesian models as well. We adhere to the concept of reproducible research. The BLFDR and BaFDR are implemented in MATLAB, and m-files with examples can be found at

`http://www.isye.gatech.edu/~brani/wavelets.html`

under BLFDR and BaFDR.

Function	n	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	512	0.4114	0.3012	0.2527	0.2130
	1024	0.2687	0.2098	0.1460	0.1099
	2048	0.1322	0.1025	0.0743	0.0573
Bumps	512	1.0058	0.9367	0.9541	1.0605
	1024	0.4902	0.5479	0.5768	0.5408
	2048	0.2785	0.3130	0.3117	0.3028
Doppler	512	0.3875	0.3657	0.3940	0.4091
	1024	0.1842	0.2085	0.2489	0.2938
	2048	0.0814	0.1031	0.1260	0.1464
Heavisine	512	0.1078	0.1440	0.1966	0.3213
	1024	0.0617	0.0873	0.1098	0.1556
	2048	0.0391	0.0609	0.0720	0.0989

Table 18: Performance of the BaFDR. The average MSE for 1000 simulations with $\alpha = 0.05$ and $\pi_0 = 0.90$ `coarsest`=5 for all.

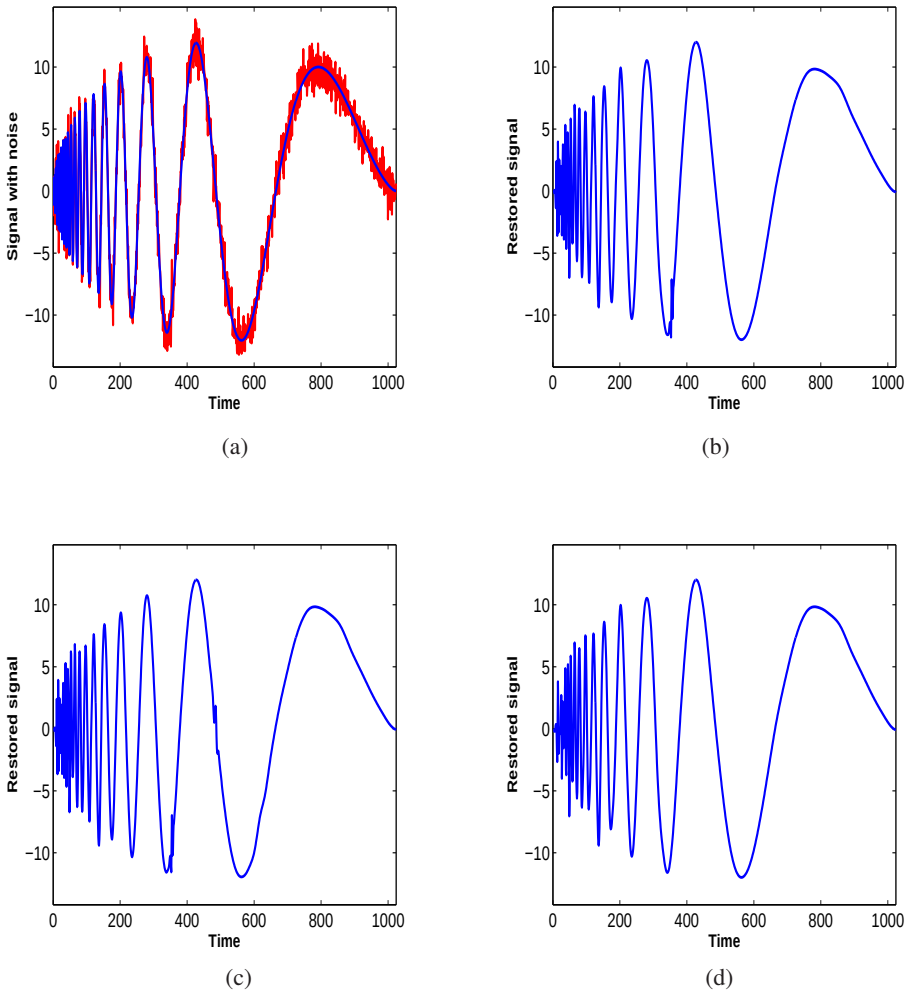


Figure 10: (a) Doppler signal with noise (SNR=7); (b) BLFDR with $p_0 = 0.95$; (c) BLFDR with levelwise p_0; and (d) BaFDR with $\alpha = 0.05$.

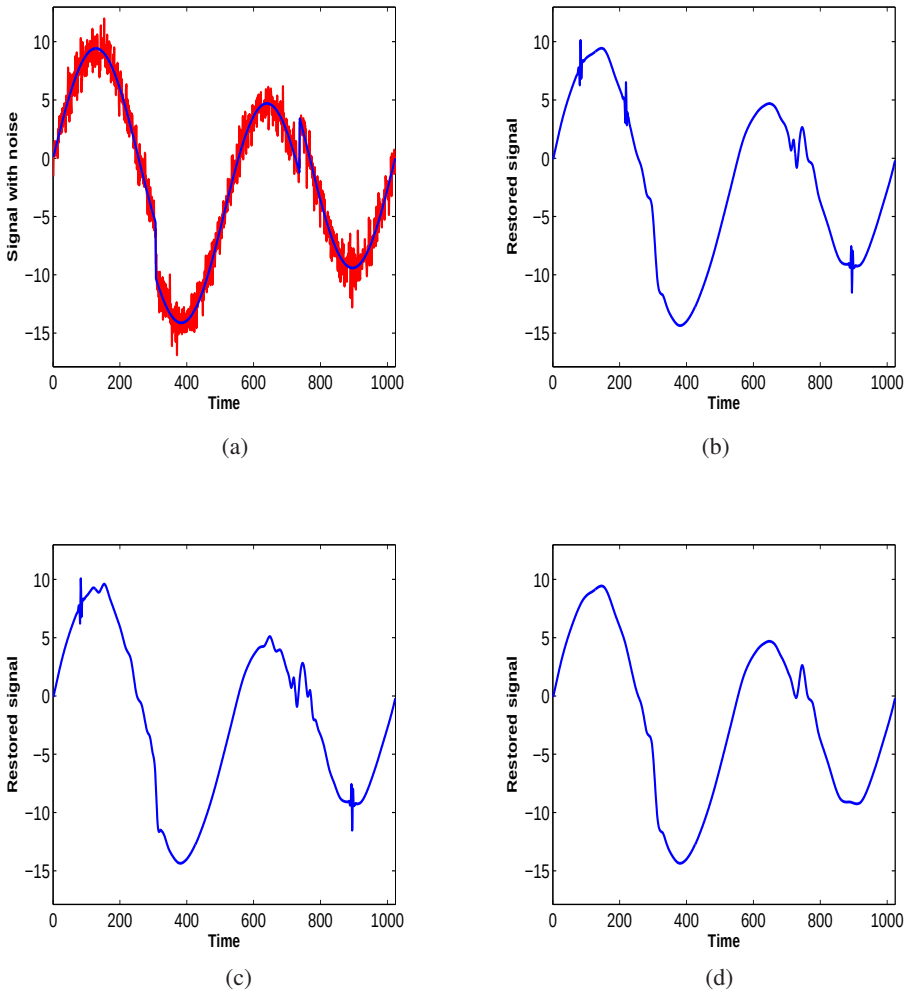
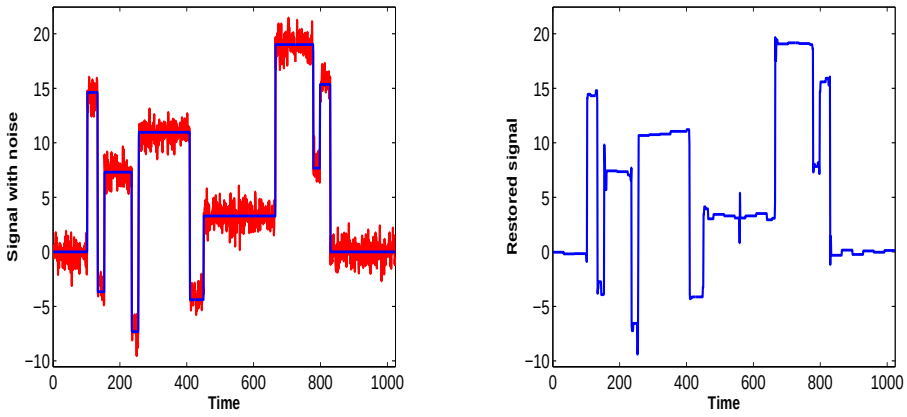
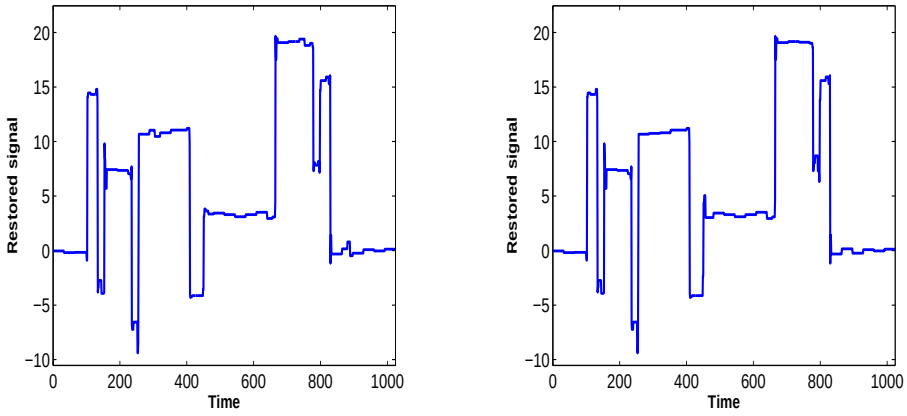


Figure 11: (a) HeavySine signal with noise (SNR=7); (b) BLFDR with $p_0 = 0.95$; (c) BLFDR with levelwise p_0; and (d) BaFDR with $\alpha = 0.05$.

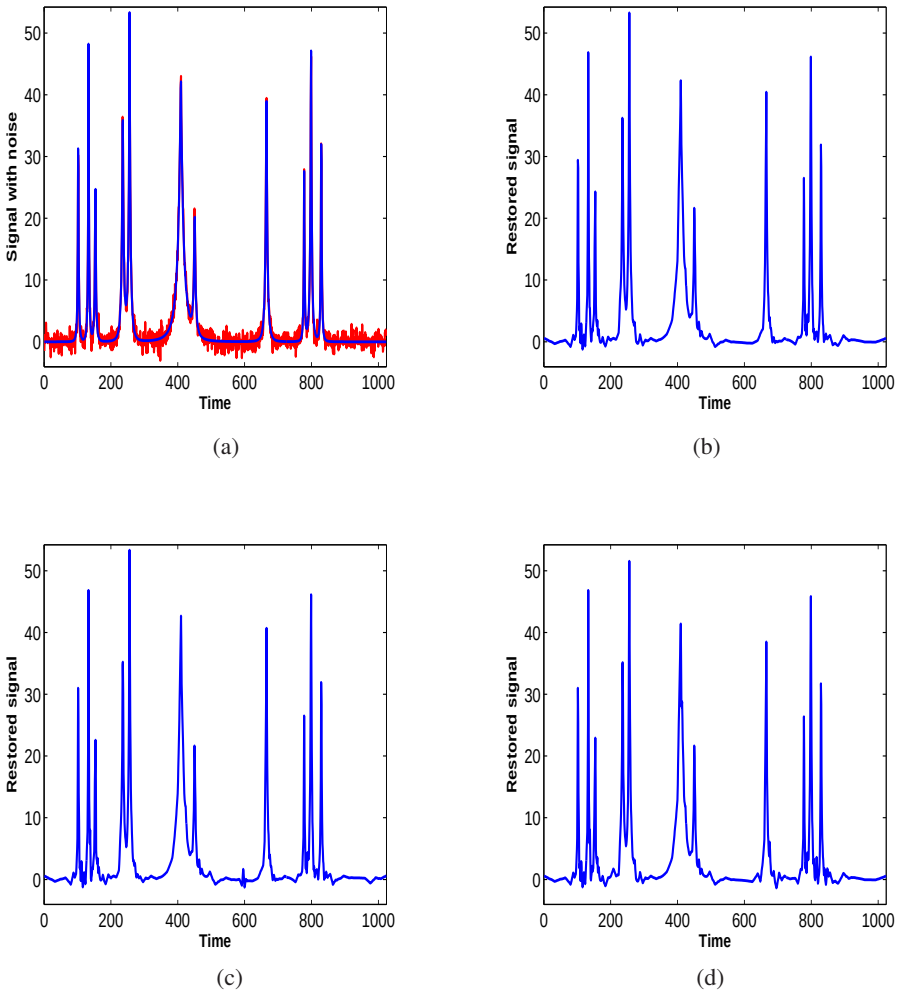


(a) (b)



(c) (d)

Figure 12: (a) Blocks signal with noise (SNR=7); (b) BLFDR with $p_0 = 0.95$; (c) BLFDR with levelwise p_0; and (d) BaFDR with $\alpha = 0.05$.



(a) (b)

(c) (d)

Figure 13: (a) Bumps signal with noise (SNR=7); (b) BLFDR with $p_0 = 0.95$; (c) BLFDR with levelwise p_0; and (d) BaFDR with $\alpha = 0.05$.

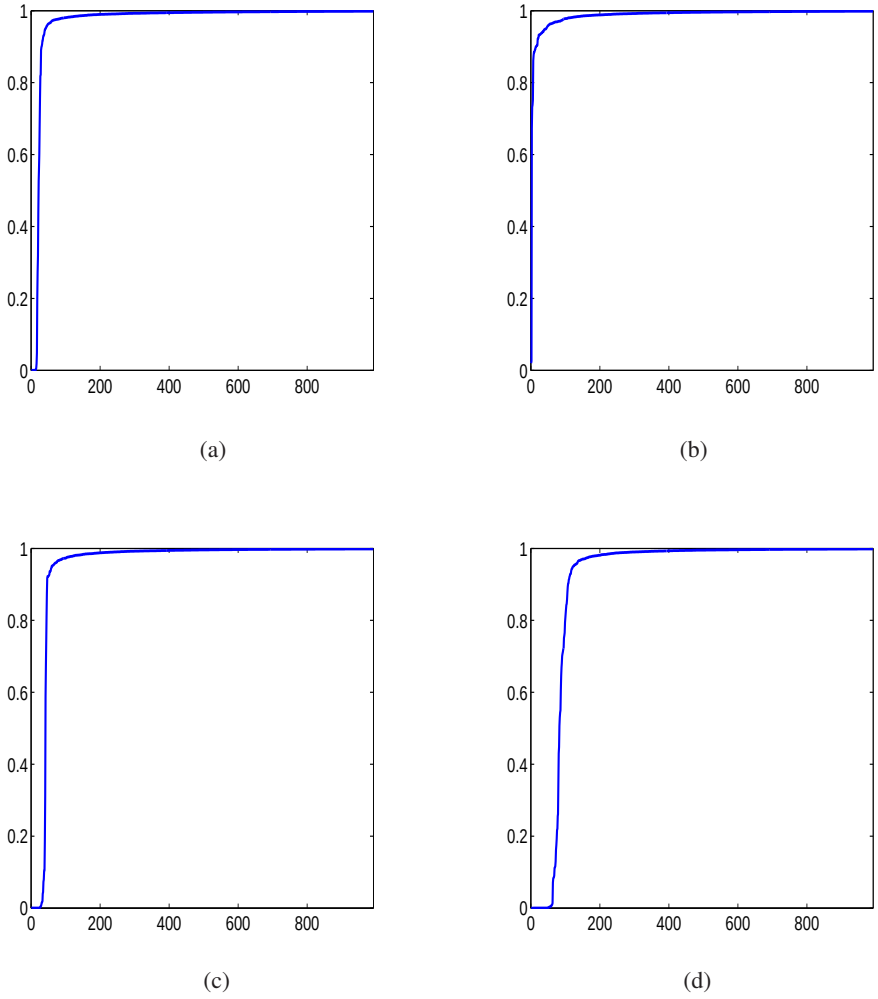


Figure 14: Ordered posterior probabilities (from BaFDR) for (a) Doppler signal, (b) Heavy-Sine, (c) Blocks, (d) Bumps.

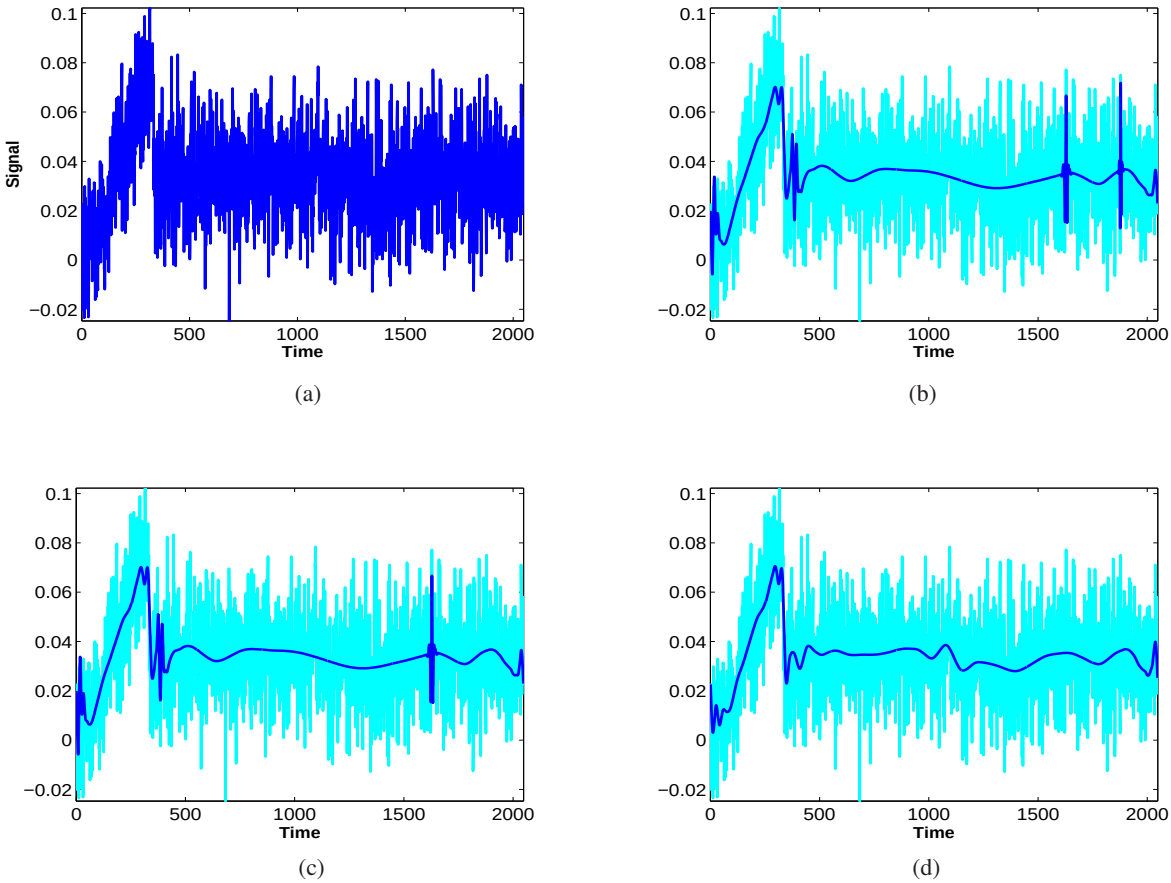


Figure 15: (a) Original AFM signal; (b) Smoothing with BaFDR; (c) Smoothing with BLFDR with $\pi_0 = 0.999$ fixed for all levels; and (d) Smoothing with BLFDR with level-dependent π_0 but fixed $\gamma = 5$.

CHAPTER IV

LINEAR FEATURE IDENTIFICATION AND INFERENCE IN NANO-SCALE IMAGES

4.1 Chapter Introduction

Nanoscale materials (i.e. materials designed on the scale of 10^{-9} meters) have been growing in interest in recent years. This is due to the emergence of nanotechnology as a field of interest in technology and to the miniaturization limitations of current technology. Nanoscale-designed materials promise to have radically different properties than their bulk counterparts. For example, the photoluminescence properties of materials change significantly in nanomaterials. Widely discussed, carbon nanotubes have been either semiconducting or metallic and have vastly improved strength over any bulk carbon.

It is important, then, to be able to characterize the materials being used in order to fully understand the properties that they exhibit. A tool crucial to this characterization and understanding is the Transmission Electron Microscope (TEM) . In order to view and understand the arrangements of atoms at an atomic scale, a high resolution transmission electron microscope is necessary. Furthermore, tools helping to analyze the images taken from the microscope could vastly enhance the ability of scientists to understand the phenomena that occur when designed at the nanoscale.

Crystalline materials are made up of atoms in specific sites within unit cells. These attributes of crystalline materials help define the many attributes that the bulk material shows. The size of these unit cells is on the order of only a couple of angstroms (10^{-10} meters) and so imaging them is somewhat of a challenge. This is solved through the use of a TEM in which resolutions up to one angstrom have been achieved. In a nanoscale world, the easy and reliable measurement of these properties of the crystals is vital to the characterization of the materials being used.

Lattice spacing determination in high resolution electron microscope images is a key way in which a material can be characterized and studied. The spacing of unit cells of atoms and the angles

that the sides of the unit cells make are both techniques in characterizing a crystalline material. Also interesting are spaces in the crystal where this regularity breaks down. This can symbolize defects in the crystal structure, such as dislocations, point defects, and planar defects. Such defects can have a large impact on the properties of the material. In many cases, it can be difficult to see the presence of nanoscale particles without an aid.

When taking a low magnification, high resolution (500kx, 1.2 angstrom resolution) images, layers of atoms manifest themselves as a series of parallel lines. The separation between these lines can be used to determine the separation between the layers of atoms. This is important in determining several important factors about the material, including the crystallographic orientation and some mechanical properties. Most of the time, these lines are visible to the human eye and currently are measured by hand with a magnifying glass, after the pictures of the specimen have been developed, and after the specimen is no longer in the microscope. More and more, these microscopes have digital cameras installed on them, so the ability to make these measurements immediately, while the specimen is still in the microscope is an extremely useful tool to researchers. Knowing what you have already measured while you are still working on the microscope can lead to better analysis and an easier time of making all of the correct measurements. Moreover the visual scans are not very accurate and often miss hidden crystallographic orientations. Therefore, developing a tool to automate the process of determining the spacing and orientation of the lattice of atoms could be important to the development of the understanding of materials and their properties at the nanoscale.

Figure 16, shows the typical TEM image of the ZnS structure. The “linear” structure (parallel lines of different orientation) formed by an atomic lattice is clearly visible. The difference in orientation may come from the following sources: (i) different materials will have different orientation; (ii) often the layer that is below the surface can be seen, and this creates the additional orientation and (iii) different areas of the crystal can be oriented differently. Our primary goal is to detect this linear structure, or more specifically, to find the parallel lines, their relative orientations, and distances between parallel lines of the same direction. All this information is useful in the following important applications. First of all, by knowing of the relative orientation of different materials we can learn more about the crystallographic structure of the interfaces of the materials. Second,

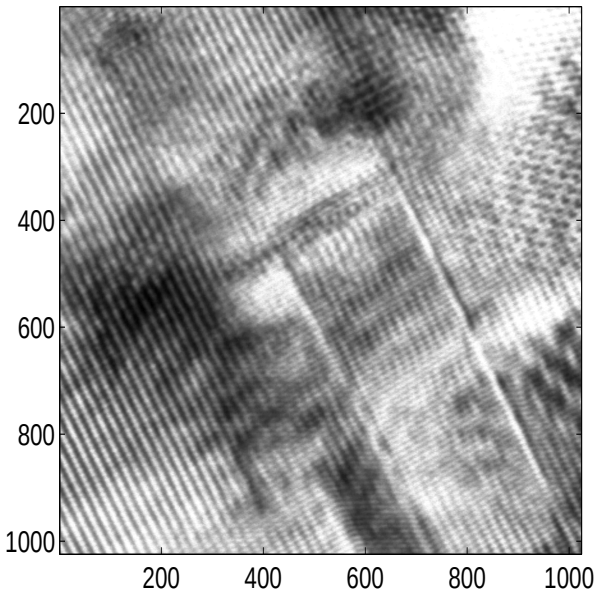


Figure 16: Example of the TEM image of the ZnS structure. Parallel lines formed by an atomic lattice are clearly visible. They form approximately 30° angle with the y-axis.

the knowledge of the orientation of the surface layer and the layers below it, coupled with diffraction pattern and images at higher resolution, can give us signature of the material, its structure and properties. Finally by learning the distances between parallel lines the distance between atoms, the lattice spacings can be determined.

4.2 The Hough Transform

The straight lines are the pattern of interest in the images. There are several different ways of representing the straight line in $\mathbb{R}^2$. For convenience, the *normal* representation of the line is used:

$$x \cos\theta + y \sin\theta = \rho,$$

where ρ is the length of a normal from origin to the line and θ is the angle of the orientation of ρ with respect to the x-axis (See Figure 17). Simple geometry shows that the angle between the line and the "negative" y-axis is also θ. In the future, we will refer to an orientation as an angle formed by the line and "negative" y-axis.

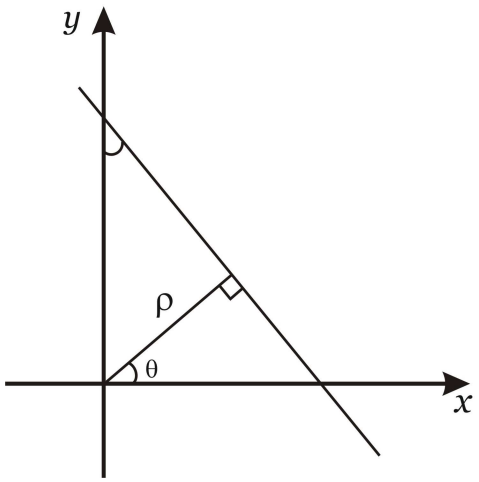


Figure 17: Normal representation of a straight line in $\mathbb{R}^2$.

There are many methods in image processing for the detection of lines. One of the most popular is the Hough Transform (HT) .

The Hough transform (HT) is a well known tool for the detection of the straight lines. Paul Hough [37], deduced the method in order to detect the straight line tracks left by the charged particles in a bubble chamber. His proposal was based more on intuition than on formal mathematical ground. Later, Duda [24], introduced the (ρ, θ) parametrization, and Deans [20] showed that the Hough transformation is in fact a special case of the well known Radon transform.

The Hough transform involves three main steps. The first step is the computation of a binary edge image $I(x, y)$. The edge description is commonly obtained from a feature detection methods such as the Laplacian of Gaussian method, the zero-crossing method, Robers Cross, Sobel, or Canny edge detector, and it is usually noisy, i.e. it contains multiple edge fragments corresponding to a single whole feature.

The second step is the evaluation of the formula

$$HT(\rho, \theta) = \iint_{\mathbb{R}^2} I(x, y)\delta(\rho - x\cos\theta - y\sin\theta)dxdy, \tag{43}$$

where δ is the standard delta function, $\delta(x) = 0 \quad \text{for all} \quad x \neq 0$. Equation (43) is the mathematical representation of the standard Hough transform. Any grayscale image is stored in computer as a matrix. Thus function $I(x, y)$ in the Equation (43) will be discrete. This requires the discretization of the the parameters of the lines ρ and θ. Hence the integrals in (43) will be represented as sums.

For an $N \times N$ image, discrete values of the (ρ, θ) variables, within the intervals $[-N/\sqrt{2}, N/\sqrt{2}]$ and $[0, \pi]$ respectively. Discretization of the parameter θ may, for example may be from 1 to 180 degrees in steps of $\Delta\theta = 1$, or one may choose half-degree step $\Delta\theta = 0.5$, to increase sensitivity. One can discretize parameter ρ in similar way with different values for the step $\Delta\rho$. The size of the steps creates the dimensions of a "probe line" or rectangular window/band along which the formula is evaluated. Simply stated, the Hough transform computes the sum of the edge map I, along the straight "probe lines" defined by the polar parameters (ρ_n, θ_m), and stores the values in the corresponding bins $HT(\rho_n, \theta_m)$ forming the accumulator matrix R. The Hough transform could be generalized by changing the argument of the delta function. A *generalized* Hough transform can be used for the detection of regular curves such as circles, ellipses, *etc*, and it is can be employed in applications where a simple analytic description of features of a pattern of interest is not possible.

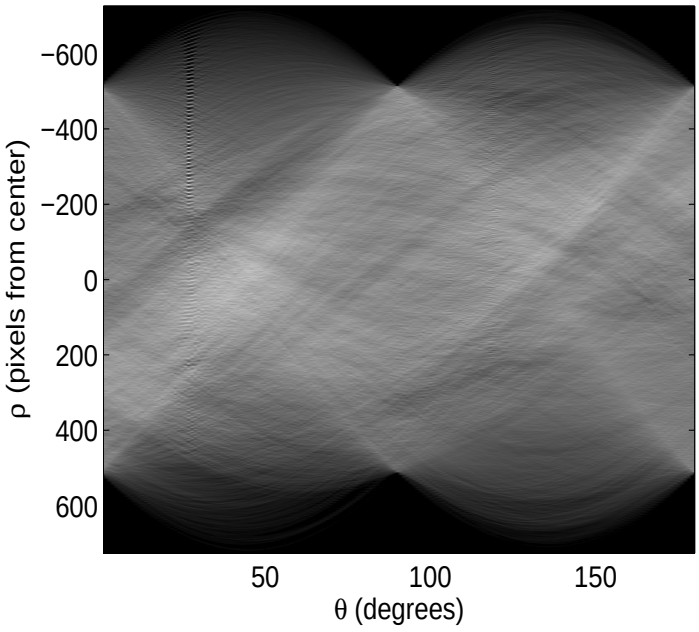


Figure 18: Example of the HT applied to image in Fig 16. Notice irregularities at 27°. These correspond to parallel lines formed by an atomic lattice clearly visible in Fig. 16.

The last step is the analysis of the output. There are number of methods which one can employ in order to extract bright points (*local maxima*), from the accumulator, in other words, unique (ρ, θ) points corresponding to each of the straight lines in the image. The simplest method is the *relative thresholding*. One could take only those local maxima in the accumulator whose values exceed

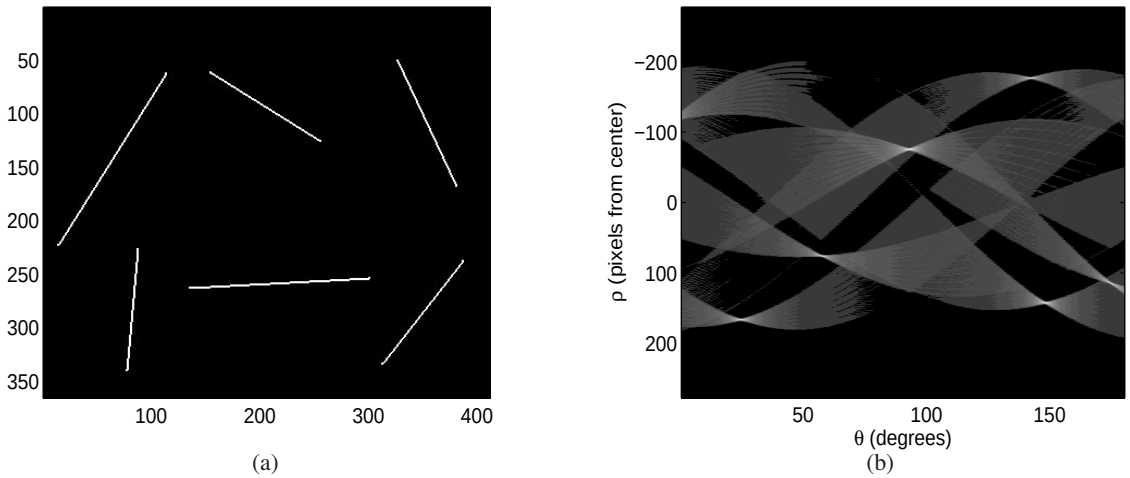


Figure 19: An image with straight lines and the HT of this image. Notice a distinct "butterfly" shape formed by the lines after the HT. The line in the neighborhood of the point $(250, 200)$ which forms an angle slightly over 90° with the y-axis in a) will correspond to the "butterfly" with its center approximately at $(-90, 90)$ in b).

some fixed percentile. However, there are many local maxima that do not correspond to straight lines because of sensitivity of the Hough transform to the correlated noise. Hence, the relative thresholding generally performs poorly.

Figure 19 shows the representation of straight lines in the Hough transform output. The lines have a more complex representation than just the local maxima. One can clearly see the distinct distribution of intensity associated with each straight line featured in image space. The distribution has the appearance of a butterfly with its wings extended in the θ direction. Therefore, instead of looking for local maxima one can be looking for this particular distribution around local maxima. This can be done by using a mask or filter that matches the distribution under investigation. The analytical form of the "butterfly" distribution in transformed space has been deduced using a step by step geometric approach and a limiting process. If the line under detection has a normal which subtends an angle α with the x-axis, then

$$HT(\rho, \theta) = \frac{1}{|\sin(\theta - \alpha)|}.$$

More information about the Hough transform, butterfly distribution and filtering, and other Hough transform techniques can be found in Leavers [42] and Illigworth and Kittler [38].

4.3 Method Description

As mentioned above, the first step is the creation of the binary edge map, edge detection. The Canny method of edge detection was found to be particularly well suited for this purpose. The method finds edges by looking for local maxima of the gradient of the image, calculated using the derivative of a Gaussian filter. The method uses two different thresholds. The first threshold detects strong and weak edges, and the other includes the weak edges in the output only if they are connected to strong edges. Compared to other standard methods, the Canny algorithm is less likely to be "fooled" by noise, and is more likely to detect true weak edges.

After the edge detection is complete, the standard Hough transform is performed to obtain the accumulator matrix R. The parallel lines of different orientation formed by the atomic lattice is pattern of interest. Due to the physical nature of the images, the presence of the parallel lines throughout the whole image is expected. Parallel lines with the same orientation will be represented as bright points at one specific column (equivalently, angle) of the accumulator matrix R. Thus it is expected that some angles would have more energy than that of others. The energy function Ang, can be obtained from the accumulator matrix R, as follows

$$Ang(j) = \sum_i R_{i,j}^2, \quad \text{for} \quad j = 1, 2, \ldots 180.$$

For convenience, Ang is normalized, so that it has a zero mean, and a unit sample variance. Figure 20 illustrates the application of the above concept to the image in Figure 16.

The energy function helps identify the angles that correspond to structured patterns of parallel lines. These angles will be represented on the graph of the energy function as sharp peaks. On the other hand, those that do not correspond to such patterns, for example some instances of correlated noise, will show up as flatter or less sharp local maxima. This behavior is captured in Figure 20, where three distinct peaks can be observed at 27°, 90°, and 119° elucidating patterns of parallel lines aligned at those angles. There is also a flat local maxima present around 45° which does not correspond to a pattern of interest. In this way, the identification of peaks in the graph of the energy would accurately determine the orientations of the parallel lines formed by the atomic lattice.

Because of their localization property, wavelets are employed as an appropriate tool for detecting peaks. To this end, the non-decimated wavelet transform of the energy function Ang is performed

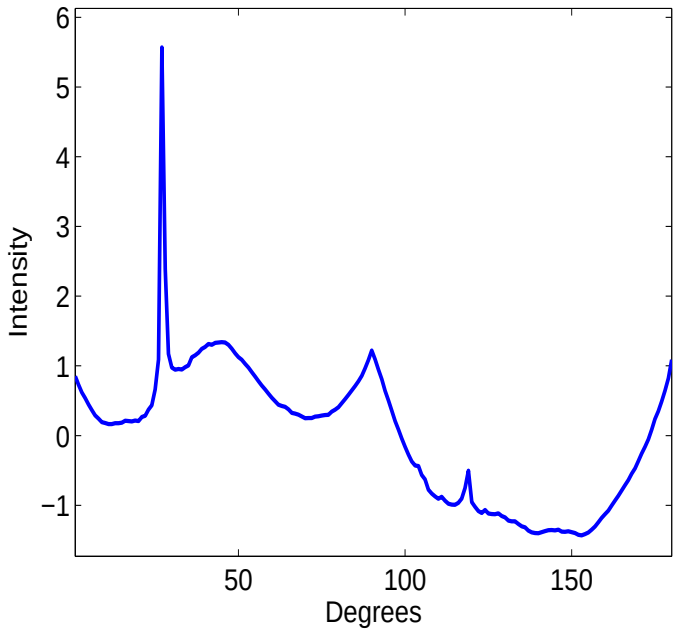


Figure 20: Energy plot Ang of the HT accumulator matrix R for the image in Fig. 16. Notice three distinct peaks at $27°$, $90°$ and $119°$. The peak at $27°$ corresponds to a major visible orientation of the image in Fig. 16. The peak at $119°$ corresponds to a second orientation, which is barely visible.

using the Haar wavelet. Using two levels of the decomposition has proved to be sufficient to detect the peaks. The coefficients at the two detail levels with large absolute values would correspond to irregularities in the energy function Ang since the levels are close to the first and second discrete derivatives of the function. For the Haar wavelet, if properly scaled, the first two levels are exactly the first two numerical derivatives. The following rule is used in order to determine the significance of an angle. The angle i, for $i = 1, 2, \ldots 180$, is considered to be a significant if it satisfies:

$$\alpha d_{i,1} + (1-\alpha) d_{i,2} > \alpha q_{p_1}(\text{level1}) + (1-\alpha) q_{p_2}(\text{level2})$$

and

$$d_{i,1} > 0, \quad d_{i+1,1} < 0,$$

where $d_{i,1}$ and $d_{i,2}$ are the coefficients of the first and second levels of the decomposition, respectively, corresponding to a given angle i. The $q_p(\text{level k})$ is the $p \cdot 100\%$ quantile of the coefficients of level k, for $k = 1, 2$. The relative importance of the levels is selected by the value of α. The

default values of p_1, p_2, and α are 0.90, 0.75, and 0.75, respectively. These default parameters provide very good results in noisy and real life images. These parameters can be changed in order to increase sensitivity and detect otherwise overlooked features. The first expression in this rule finds all significant coefficients in the decomposition which correspond to irregularities in the function Ang such as fast decay, discontinuity jumps, peaks, *etc*. The second guaranties that the suspicious angle is local maxima. This rule proves to be a very efficient in finding peaks of the function Ang.

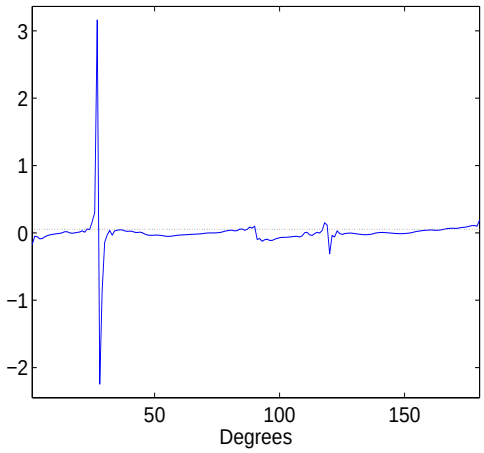


Figure 21: Coefficients of the first level of non-decimated wavelet decomposition of the function Ang in Fig 20. Notice the behavior of the coefficients at $27°$, $90°$ and $119°$.

After analyzing real life images, an interesting feature is found: the peak at $90°$ almost always appears in the graph of the energy function Ang. It can be shown that this peak is an construct of the Hough transform. For instance, if Bernoulli random noise is used to create an original edge map image (with probability p for a given pixel to be 1), then the energy function Ang will have a distinct shape shown in Figure 22.

The sharpness of the peak depends on the percentage of ones in an edged map. Sometimes, some parts or the even a whole image will not posses any features of interest; still their energy function will have a distinctive shape as in Figure 22. Thus, the shape of the energy function of an image with only noise, can be used as a template function for the absence of linear structure in images. It is very unlikely that a realistic nanoscale image would have parallel lines at $90°$. Hence, peaks at $90°$ can be ignored.

After determining the orientation of the parallel lines, the next step is to find their location by

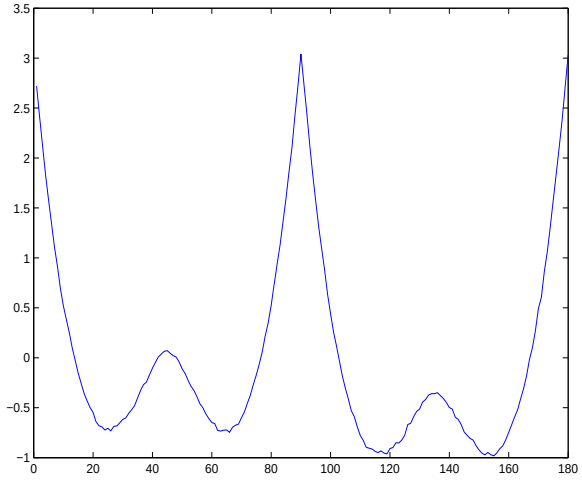


Figure 22: Function Ang that corresponds to the $0-1$ image with probability 0.1 for a given pixel to be 1.

analyzing the columns of the accumulator matrix R. The angles, which were found in the previous step, would determine exactly which columns of the accumulator matrix are needed to be analyzed. For example, in the image in Figure 16 only two orientations were detected – one at 27° and a second at 119°. Therefore, one should focus on only the 27^{th} and 119^{th} columns of the matrix R. Figure 23 shows the 27^{th} column of the accumulator matrix R.

The relative thresholding will correspond to a horizontal cut on the graph in Figure 23. If the value of the threshold is too large (cut high) then too few lines would be detected. Lines that are close to the end of the image would be ignored, as well as the lines whose length is relatively small. If the value of the threshold is too small (cut low), then too many "noise lines" would be detected which would be useless for the analysis. To this end, a compromising thresholding technique based on wavelets has been developed.

Let dis denote the column of interest of the accumulator matrix R. The length n of dis depends on the size of the original image. The signal dis is decomposed using a wavelet transform. The Vaidyanthan wavelet filter was selected for the decomposition. The sound-like form of the signals gives a strong indication that the Vaidyanathan wavelet is appropriate, since this filter has been optimized for speech coding. The number of levels in the wavelet decomposition is selected as

$$n_l = \left\lfloor \frac{\log_2(n-1)}{2} \right\rfloor . \tag{44}$$

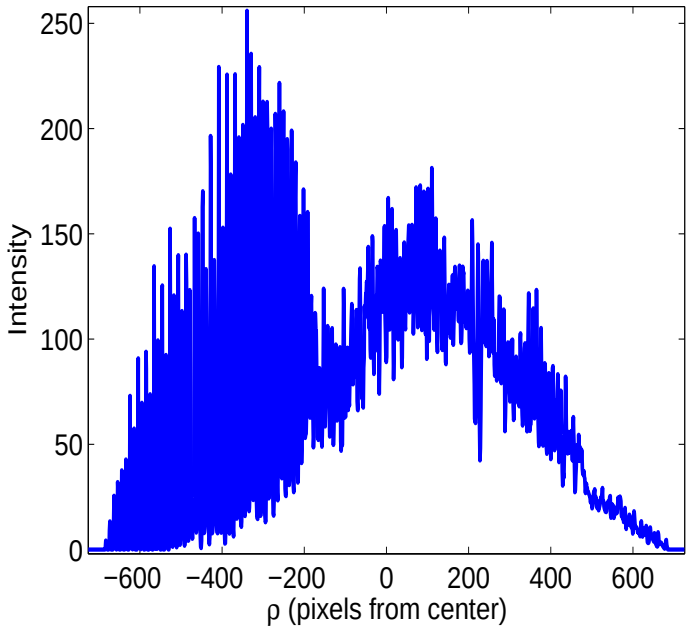


Figure 23: Plot of the 27^{th} column of the accumulator matrix R.

The wavelet decomposition is thus composed of n_l levels containing the details of the signal dis, plus the smooth part, which contains the information about the general behavior of the signal. The number n_l as in the (44), ensures that the smooth part does not contain any high frequency features. Using only this smooth part of the wavelet decomposition, one can create a flexible threshold. Application of the inverse wavelet transform to the smooth part only produces the smoothed version of the signal dis. This reconstruction is used as a threshold criteria. The dotted curve in Figure 24 a) shows the smoothed signal dis. After shifting the restored signal by an appropriate constant, one can consider everything below the curve as insignificant. The standard deviation of the absolute values of the residuals of the signal dis and its restored smoothed representation is selected as the shift constant. The solid curve in Figure 24 a) represents shifted smoothed signal dis, which is used as a threshold. By itself this threshold is not selective enough, for it selects too many lines. The Vaidyanthan wavelet transform is then applied to the original signal dis again, but this time we ignore the first level (finest level) of the decomposition. Similarly, the inverse wavelet transform is applied to the smooth part of the decomposition. The restored signal repeats the behavior of the original signal dis almost perfectly. Figure 24 b) shows the restored signal with only one level of

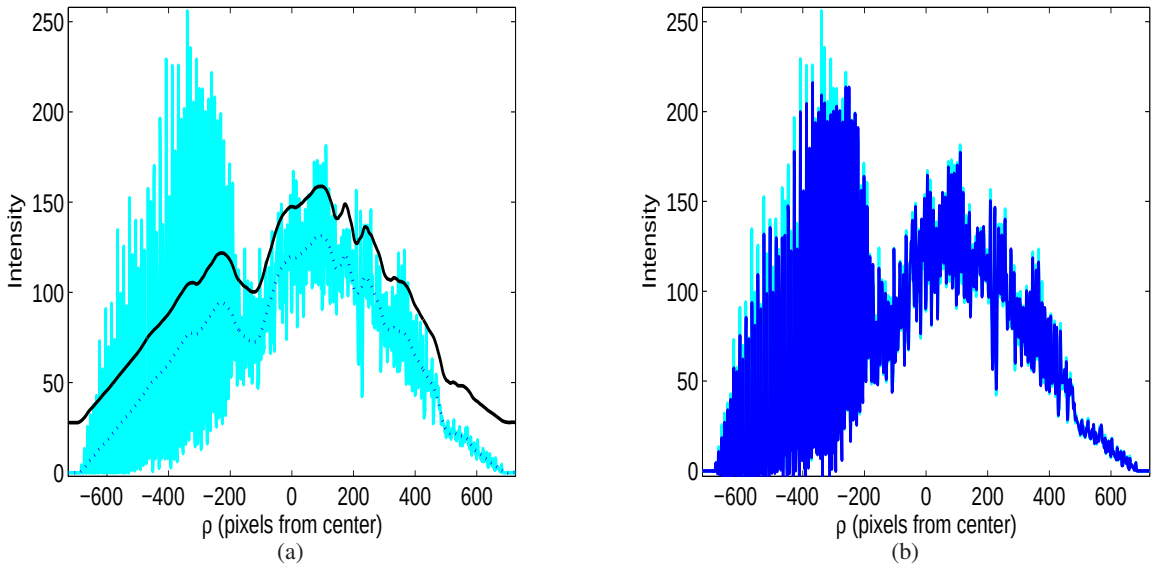


Figure 24: (a) Cyan - original signal. Dotted line - smoothed signal dis. Solid curve - shifted smoothed signal. (b) Cyan - original signal. Blue - restored signal without finest level.

decomposition. There are key differences in a restored signal and the original signal which make it easier to eliminate most of the extra lines. The restored signal is based on only half of the data points of the original signal. It averages values of neighboring data points. The restored signal always shrinks towards its average. Everything below the restored signal is ignored. The combined two-step thresholding produces good results.

Another challenge comes from the pixel representation of a straight line. The granularity of the pixel representation of a line can be coarse enough for the Hough transform to detect two or more lines of the same orientation, where, in fact only one line exists (see Figure 25). Some orientations tend to create more neighboring parallel lines than the others. Experiments with different images have shown that orientations which favor generation of several lines are located in the neighborhood of the local minima of the function in Figure 22. This creates a problem for the analysis of the lattice spacing – the distance between parallel lines of the same direction. The introduction of several extra lines in close proximity will act as noise.

When working with real life images with visible linear structure, continuous straight lines are rarely found, since lines are broken into pieces. The lines are not always straight, due to the defects

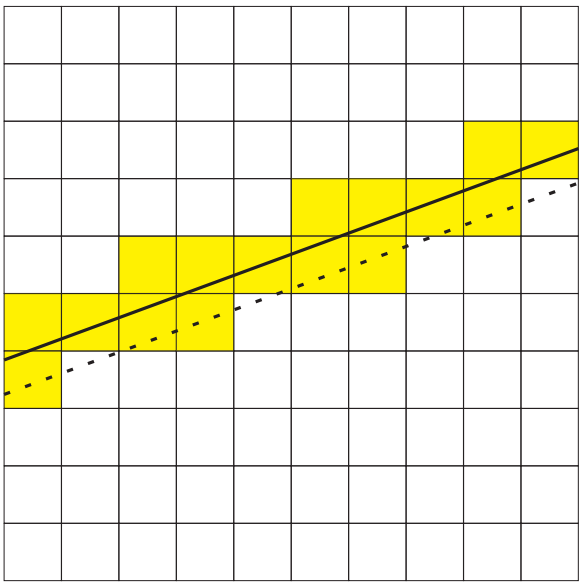
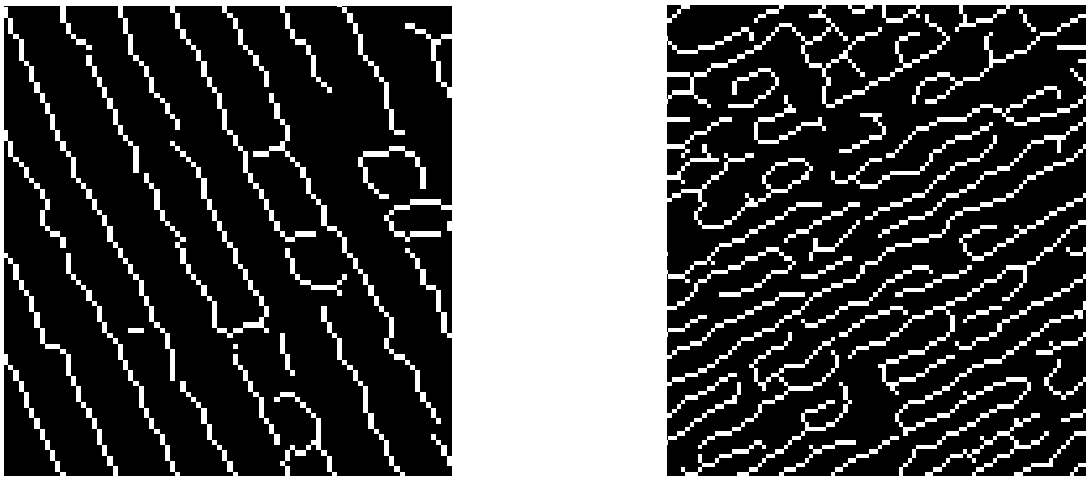


Figure 25: Theoretical line(solid); its pixel representation. The dotted lines represents a neighboring line, which might get detected since it goes through many pixels of the pixel representation.

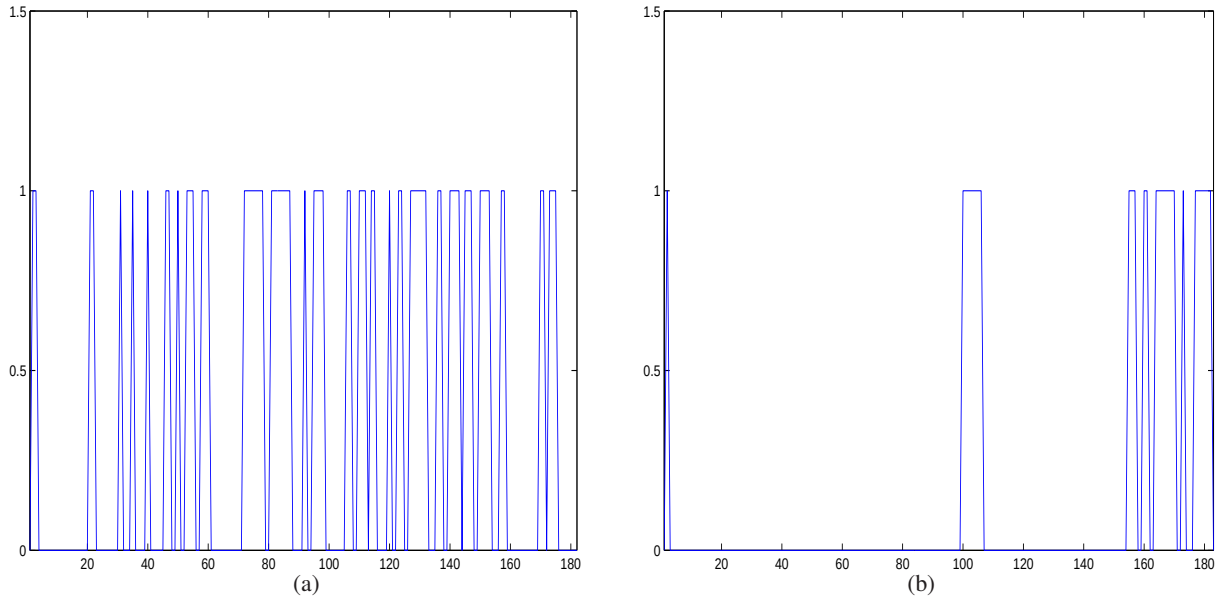
in a materials, and/or poor image quality. Thus, after edge detection there would be a noisy image with fragmented lines. Even in the continuous fragments of the same line one can observe small shifts (see Figure. 26). All these factors cause the single line to create several extra lines, which brings more noise into the analysis.

The following procedure has been developed in order to solve this problem. Among detected lines, those that are one, two or three pixels apart were singled out. From the equations of these lines, pixel representations are generated. These representations are the coordinates of pixels outlining the original lines. Using these pixel coordinates the signal of zeros and ones can be obtained from the edge map, which would correspond to a line under investigation. One could choose several measurements in order to decide if the line is significant: the simple count of ones, the maximum length of the continuous run of ones, or the average length of runs of ones. The significance measure could help to differentiate between the actual line and the neighboring noise lines. Figure 27 illustrates the representation of a detected line and its neighbor. Clearly the second line b) is only detected because of the "spill-over" from line a). Therefore it could be ignored. This step improves the results significantly.



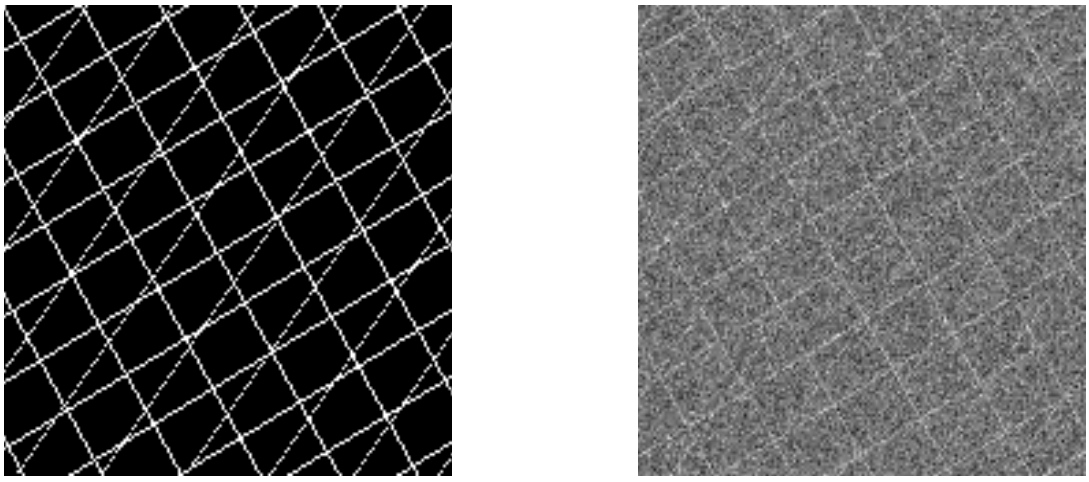
(a) (b)

Figure 26: A close-up on the linear structure of the ZnS structure. Notice how the lines formed by an atomic lattice are not exactly straight or continuous. One can clearly see that the lines are fragmented and continuous segments contain small shifts. All this generates problems in the analysis.



(a) (b)

Figure 27: Representation of the two neighboring lines detected after thresholding. Line (b) is clearly detected as a "spill-over" from line (a) and can be ignored.



(a) (b)

Figure 28: (a) Example of the deterministic image with three sets of parallel lines; (b) Example of the deterministic image with added normal noise (SNR=0.5). Notice how the lines with 144° orientation are barely visible in the noise.

4.4 Results

4.4.1 Deterministic Images

Several synthetic images of size 1024×1024 have been created in order to test performance of the method. All images consist of fifty parallel lines of various orientation with a fixed distance between them. Different levels of normal noise have been added to each image. Table 19 shows the results of the analysis of these images.

Notation: "# of lines" refers to number of lines detected. "Distance" denotes estimated average distance detected. "Std" refers to standard deviation of the estimated distance between the lines.

Since all the lines have a fixed distance between them, standard deviation theoretically should be zero. However, because of approximations in constructing lines (i.e. pixel representation) some small variability is expected. Images 1 through 4 have single orientation, where images 5-7 have two sets of parallel lines and image 8 has three. Figure 28 shows an enlarged portion of image 8 together with the noise added to this image.

4.4.2 Nanoscale Images

Application of the method is demonstrated on the two TEM images of the ZnS structures. Figures 29 and 30 show the images. Because of their large size the images have been broken into 4 subimages in the following manner

(1, 1)	(1, 2)
(2, 1)	(2,2)

. Each subimage is analyzed separately. The results for each image can be found in Tables 20-22. The results are given for the default parameter settings, using three-pixel difference elimination for close lines. A 10% trimming from above is applied, for the estimation of the average distance between parallel lines or lattice spacing. The theoretical lattice spacing for the ZnS materials is $1.66 \cdot 10^{-10}$ m. This if for the inside of material in ideal condition, as on the surface the spacing may vary. As one can see from the Tables 20-22, the estimate of the major layer is almost always greater than the theoretical distance. However, the estimate of distances for visible layers under the surface is very close to the theoretical distance.

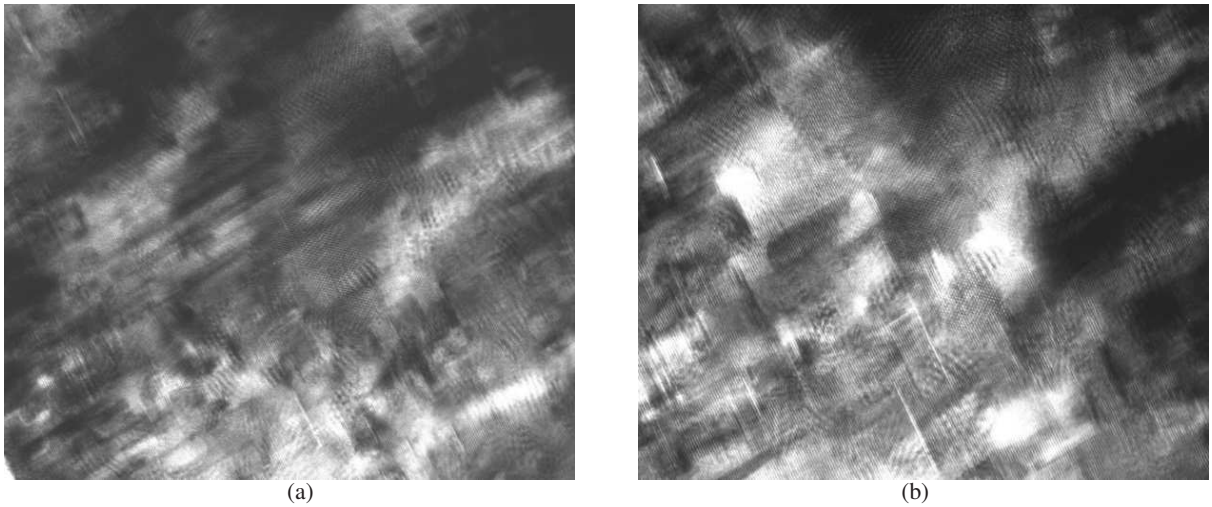
On the surface, the atoms experience different energetic responses than in the bulk of the material. In order to compensate, the atomic layer relaxes (spreads out) or otherwise rearranges itself. This is the major reason why the estimated distance of the main visible layer exceeds the theoretical distance. For some of the materials the way atoms rearrange themselves on the surface is well studied, and the lattice spacing could be determined theoretically. For the others there is no method that would give good understanding of the rearrangements and provide theoretical lattice spacing. Our method allows the experimenter to estimate the lattice spacings quite accurately.

4.4.3 Analysis of Image 1

Size 4050×5220, scanned at 2400dpi and with microscope magnification of 500,000. The image is broken into 4(2x2) overlapping images 2048×2048, which give almost complete coverage of the original image. Quality of the subimage (1, 2) is poor. No features are found with default parameter settings. Results of the analysis are in Table 20. 29° is the major visible orientation present almost throughout the entire image. 78° is visible only in part of the image.

4.4.4 Analysis of Image 2

Size 3360×4560, scanned at 2400dpi and with microscope magnification of 500,000. The image is been broken into 4(2x2) overlapping images 2048×2048, which give almost complete coverage of



(a) (b)

Figure 29: (a) Image 1: size 4050×5220; (b) Image 2: size 3360×4560, ZnS structure, microscope magnification 500,000.

the original image. 28° is the major visible orientation present almost throughout the entire image. 118° is the secondary orientation present almost always throughout the whole image, but is barely visible, and with default parameter it is detected only once. Results of the analysis are in Table 21.

4.4.5 Analysis of Image 3

Size 2415×2745, 2400dpi and with microscope magnification of 500,000 times. The image is broken into 4(2x2) overlapping images 2048×2048, which give complete coverage of the original image. Results of the analysis are in Table 22.

4.5 Conclusions and Discussion

In this chapter we propose a new method for the analysis of the nanoscale images. The proposed method produces good results in both synthetic and real nano-scale images. It is recommended to have two images of the same sample: one the rotated version of the other. Comparison of the results of both images would allow for the removal of the uncertainty that comes from the sensitivity of the Hough transform to certain orientations. This would increase accuracy in the analysis of the lattice spacing. Current 'algorithmic' methods of image rotation would preserve the image's structure completely, and the energy function would shift according to the angle of rotation (see Figure 31). This is why it is recommended to have two images.

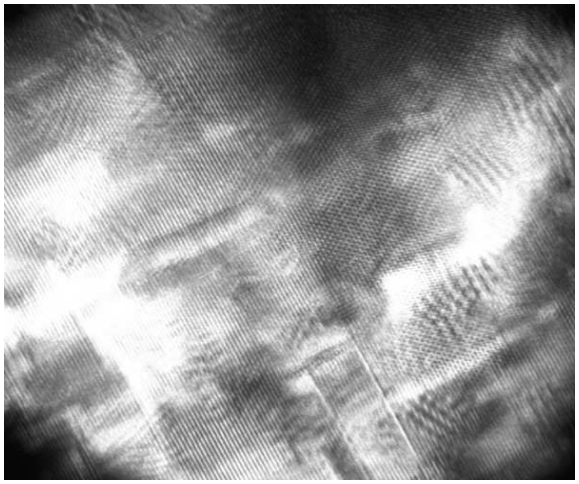


Figure 30: Image 3: ZnS structure, size 2415×2745, microscope magnification 500,000.

It is possible to perform continuous directional wavelet transform of the images with the directions defined by the output of the analysis. The continuous directional wavelet transform would extract the features of the images which are aligned in the given direction. Figure 32 shows the continuous directional wavelet transform of the image in Figure 16 in two directions 27° and 119°. The results of the continuous directional wavelet transform could tell a little bit more about structure and defects in the material for specific orientation.

We support David Donoho's initiative for reproducible research. MATLAB toolbox, tutorial file, sample images, and m-files used to produce the calculations and pictures in this chapter are available at Jacket's Wavelets page `http://www.isye.gatech.edu/~brani/wavelet.html`.

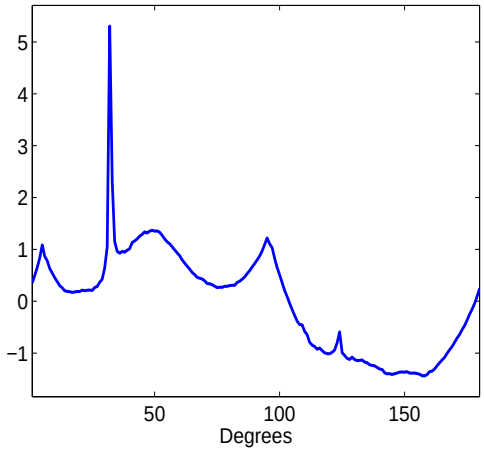


Figure 31: Plot of the energy function of the rotated by 5° image. The energy function of the rotated image almost completely preserves structure of the energy function for the original image. Everything is shifted by 5°.

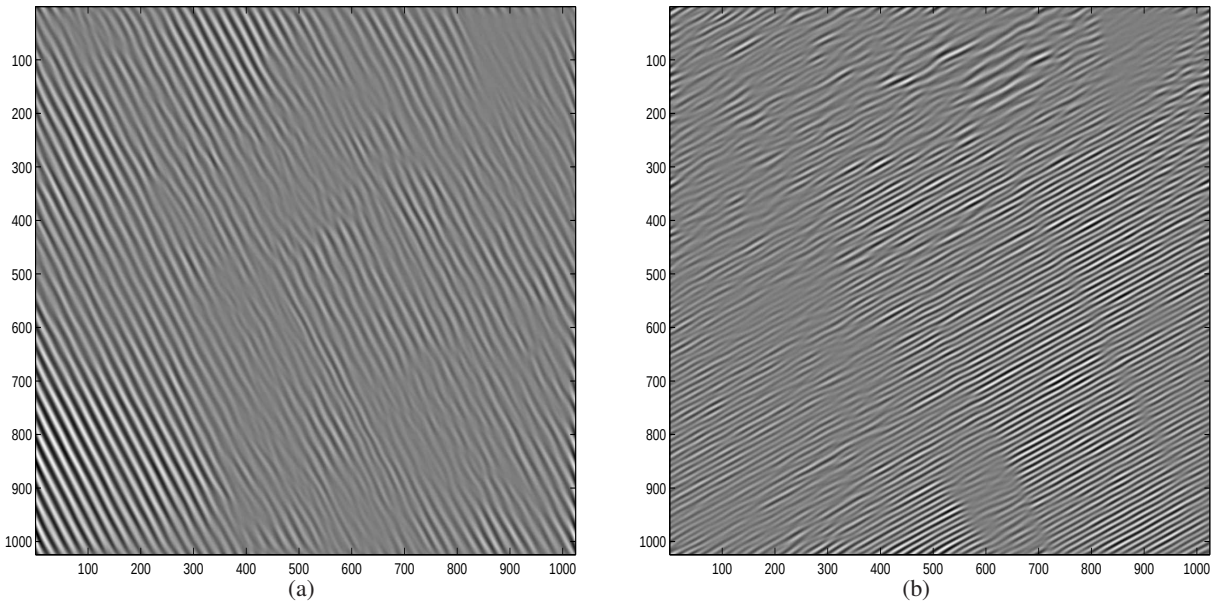


Figure 32: Continuous Direction Wavelet Transforms of image in Fig. 16 a) 27° b) 119°.

Description	SNR	0.3	0.5	0.7	1	3	No Noise
Image 1(1 set)	# of lines	51	49	49	49	49	49
Distance between the lines is 25 pixels	Distance	23.5	25.06	24.93	25	25	25
Angle 30° between lines and y axis	Std	4.445	1.31	1.31	0.87	5.8E-15	5.8E-15
Image 2(1 set)	# of lines	66	47	47	47	49	49
Distance between the lines is 20 pixels	Distance	16.18	20.06	20	20	19.58	20
Angle 150° between lines and y axis	Std	10.2	2.4	2	2	2.5	1.6E-14
Image 3(1 set)	# of lines	49	49	48	48	48	48
Distance between the lines is 20 pixels	Distance	19.58	20	20	20	20	20
Angle 120° between lines and y axis	Std	3.03	1.23	0.437	1.25	8.3E-15	8.3E-15
Image 4(1 set)	# of lines	67	50	48	46	47	48
Distance between the lines is 25 pixels	Distance	17.42	24.42	25	25	24.93	25
Angle 144° between lines and y axis	Std	8.29	5.26	1.37	1.73	1.18	1.6E-15
Image 5(2 sets) Set 1:	# of lines	70	49	49	49	49	49
Distance between the lines is 25 pixels	Distance	17.75	25.06	24.93	25	24.93	25
Angle 30° between lines and y axis	Std	7.25	1.15	1.31	1.07	1.8	1.51
Set 2:	# of lines	117	64	50	49	49	49
Distance between the lines is 25 pixels	Distance	10.9	19	24	25.06	25	25
Angle 144° between lines and y axis	Std	6.16	7.49	4.38	1.42	0.41	0.41
Image 6(2 sets) Set 1:	# of lines	86	51	49	49	49	49
Distance between the lines is 25 pixels	Distance	14.21	24.44	25	24.93	25	25
Angle 30° between lines and y axis	Std	6.95	2.84	1.23	0.97	1.51	2.55
Set 2:	# of lines	73	49	48	48	48	49
Distance between the lines is 20 pixels	Distance	15.36	20	20	20	20	19.93
Angle 120° between lines and y axis	Std	11.3	1.75	1.39	0.625	1.97	1.57
Image 7(2 sets) Set 1:	# of lines	66	49	49	49	49	49
Distance between the lines is 20 pixels	Distance	16.09	19.93	20	20	20	20
Angle 120° between lines and y axis	Std	6.95	1.64	1.23	8.2E-15	8.2E-15	8.2E-15
Set 2:	# of lines	119	72	50	50	49	49
Distance between the lines is 25 pixels	Distance	10.4	16.95	24.53	25.04	25	24.97
Angle 144° between lines and y axis	Std	5.83	7.65	2.78	1.15	0.61	0.14
Image 8(3 sets) Set 1:	# of lines	98	52	48	49	49	49
Distance between the lines is 25 pixels	Distance	12.53	24.07	25	25	25	25.06
Angle 30° between lines and y axis	Std	6.66	3.74	1.25	1.51	1.51	2.18
Set 2:	# of lines	89	51	49	49	49	49
Distance between the lines is 20 pixels	Distance	14.02	19.54	20.06	19.93	20	20
Angle 120° between lines and y axis	Std	10.35	2.71	1.57	1.15	8.2E-15	8.2E-15
Set 3:	# of lines	Failed	107	55	50	50	51
Distance between the lines is 25 pixels	Distance		11.94	22.18	25	24.95	24.52
Angle 144° between lines and y axis	Std		6.8	5.63	1.25	0.28	2.97

Table 19: Analysis of the deterministic images. Each set of lines consists of 50 parallel lines.

Subimage	Detected Direction	Average(in m.)
(1, 1)	29	2.18E-10
(2, 1)	29	2.37E-10
(2, 2)	29	2.16E-10
	78	2.49E-10

Table 20: Image 1 analysis.

Subimage	Detected Direction	Average(in m.)
(1, 1)	28	2.03E-10
(1, 2)	28	2.03E-10
	99	2.10E-10
(2, 1)	28	2.06E-10
(2, 2)	28	2.10E-10
	118	1.71E-10

Table 21: Image 2 analysis.

Subimage	Detected Direction	Average(in m.)
(1, 1)	28	2.03E-10
(1, 2)	28	2.06E-10
	118	1.89E-10
(2, 1)	28	2.03E-10
(2, 2)	28	2.30E-10
	100	1.73E-10
	118	1.81E-10

Table 22: Image 3 analysis.

APPENDIX A

DEFINITIONS, LEMMAS, THEOREMS

Theorem A. 1 *(Theorem 29.8 p.502 in [21]). Let ψ_1, ψ_2, ... be a sequence of bounded functions with $|\psi_j(x)| \leq 1$ such that the set of all finite linear combinations of the ψ_j's*

$$\bigcup_{k=1}^{\infty} \left\{ \sum_{j=1}^{k} a_j \psi_j(x) \colon a_1, a_2, \ldots \in \mathbb{R} \right\}$$

is dense in $L_2(\mu)$ on all balls of the form $\{x \colon \|x\| \leq M\}$ for any probability measure μ. Let the coefficients $a_1^*, \ldots, a_{k_n}^*$ minimize the empirical squared error

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{k_n} a_j \psi_j(X_i) - (2Y_i - 1) \right)^2$$

under the constraint $\sum_{j=1}^{k_n} |a_j| \leq b_n$, $b_n \geq 1$. Define the generalized linear classifier g_n by

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{j=1}^{k_n} a_j^* \psi_j(x) \leq 0 \\ 1 & \text{otherwise.} \end{cases}$$

If k_n and b_n satisfy

$$k_n \to \infty,\ b_n \to \infty \text{ and } \frac{k_n b_n^4 \log(b_n)}{n} \to 0,$$

then $\mathbf{E}\{L(g_n)\} \to L^*$ for all distributions of (X, Y), that is, the rule g_n is universally consistent. If we assume additionally that $b_n^4 \log(n) = o(n)$, then g_n is strongly universally consistent.

The function $\eta(x)$ is estimated from training data D_n by some function $\eta_n(x) = \eta_n(x, D_n)$.

Corollary A. 1 *(Corollary 6.2 p.93 in [21]) If*

$$g_n(x) = \begin{cases} 0 & \text{if } \eta_n(x) \leq 1/2 \\ 1 & \text{otherwise.} \end{cases}$$

then its error probability satisfies

$$P\{g_n(X) \neq Y | \, D_n\} - L^* \leq 2\sqrt{\int_{\mathbb{R}^d} |\eta(x) - \eta_n(x)|^2 \, \mu(dx)}.$$

Let $\mathcal{F}$ be a class of real-valued functions defined on $\mathbb{R}^d$, and let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. $\mathbb{R}^d$-valued random variables. We assume that for each $f \in \mathcal{F}$, $0 \leq f(x) \leq M$ for all $x \in \mathbb{R}^d$ and some $M < \infty$.

Definition A. 1 *(Definition 29.1 p.492 in [21]) Let A be a bounded subset of $\mathbb{R}^d$. For every $\epsilon > 0$. the ℓ_1-covering number , denoted by $\mathcal{N}(\epsilon, A)$, is defined as the cardinality of the smallest finite set in $\mathbb{R}^d$ such that for every $z \in A$ there is a point $t \in \mathbb{R}^d$ in the finite set such that $(1/d)\|z - t\|_1 < \epsilon$. ($\|x\|_1 = \sum_{i=1}^d |x^{(i)}|$) denotes the ℓ_1-norm of the vector $x = (x^{(1)}, \ldots, x^{(d)})$ in $\mathbb{R}^d$.) In other words, $\mathcal{N}(\epsilon, A)$ is the smallest number of ℓ_1-balls of radius ϵd, whose union contains A. $\log \mathcal{N}(\epsilon, A)$ is often called the metric entropy of A.*

Let $z_1^n = (z_1, \ldots, z_n)$ be n fixed points in $\mathbb{R}^d$, and define the following set:

$$\mathcal{F}(z_1^n) = \{(f(z_1), \ldots, f(z_n)) : \; f \in \mathcal{F}\} \subset \mathbb{R}^d.$$

The ℓ_1-covering number of $\mathcal{F}(z_1^n)$ is $\mathcal{N}(\epsilon, \mathcal{F}(z_1^n))$. If $Z_1^n = (Z_1, \ldots, Z_n)$ is a sequence of i.i.d. random variables then $\mathcal{N}(\epsilon, \mathcal{F}(Z_1^n))$ is a random variable.

Theorem A. 2 *(Theorem 29.1 p.492 in [21]) For any n and $\epsilon > 0$,*

$$P\left\{\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}\{f(Z_1)\} \right| > \epsilon \right\} \leq 8\mathbb{E}\{\mathcal{N}(\epsilon/8, \mathcal{F}(Z_1^n))\} e^{-n\epsilon^2/(128M^2)}$$

Definition A. 2 *(Definition 12.1 p.196 in [21]) Let $\mathcal{A}$ be a collection of measurable sets. For $(z_1, \ldots, z_n) \in \{\mathbb{R}^d\}^n$, let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of different sets in*

$$\{\{z_1, \ldots, z_n\} \bigcap A : \; A \in \mathcal{A}\}.$$

The n-th shatter coefficient of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(z_1, \ldots, z_n) \in \{\mathbb{R}^d\}^n} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

That is, the shatter coefficient is the maximal number of different subsets of n point that can be picked out by the class of sets $\mathcal{A}$.

Definition A. 3 *(Definition 12.2 p.196 in [21]) Let $\mathcal{A}$ be a collection of sets with $|\mathcal{A}| \geq 2$. The largest integer $k \geq 1$ for which $s(\mathcal{A}, k) = 2^k$ is denoted by $V_{\mathcal{A}}$, and it is called the Vapnik-Chervonenkis dimension (or VC dimension) of class $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n, then by definition $V_{\mathcal{A}} = \infty$.*

Theorem A. 3 *(Theorem 13.9 p.221 in [21]) Let $\mathcal{G}$ be a finite-dimensional vector space of real functions on $\mathbb{R}^d$. The class of sets*

$$\mathcal{A} = \{\{x : \ g(x) \geq 0\} : \ g \in \mathcal{G}\}$$

has VC dimension $V_{\mathcal{A}} \leq r$, where $r = \dim(\mathcal{G})$.

Corollary A. 2 *(Corollary 29.2 p.497 in [21]) Let $\mathcal{F}$ be a class of $[0, M]$-valued functions on $\mathbb{R}^d$. For every $\epsilon > 0$ and the probability measure μ,*

$$\mathcal{N}(\epsilon, \mathcal{F}) \leq \left(\frac{4eM}{\epsilon} \log \left(\frac{2eM}{\epsilon} \right) \right)^{V_{\mathcal{F}^+}}.$$

$\mathcal{F}^+$ is defined by following:

$$\mathcal{F}^+ = \{\{(x, t) : \ t \leq f(x)\}; \ f \in \mathcal{F}\}.$$

Lemma A. 1 *(Borel-Cantelli lemma, p.585 in [21]) Let A_n, $n = 1, 2, \ldots$ be a sequence of events. Introduce notation*

$$[A_n \ i.o.] = \limsup_{n \to 0} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=1}^{\infty} A_m,$$

("i.o." stands for "infinitely often.") If

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty$$

then

$$P\{[A_n i.o.]\} = 0.$$

Theorem A. 4 *(p.585 in [21]) If for each* $\epsilon > 0$

$$\sum_{n=1}^{\infty} P\{|X_n - X| \geq \epsilon\} < \infty,$$

then $\lim_{n\to\infty} X_n = X$ *with probability one.*

Theorem A. 5 *(Cauchy root test) Let* $\sum_k a_k$ *be a series with* $a_k \geq 0$*, and let*

$$\rho \equiv \lim_{k\to\infty} a_k^{1/k}.$$

1. *If* $\rho < 1$*, the series converges.*

2. *If* $\rho > 1$ *or* $\rho = \infty$*, the series diverges.*

3. *If* $\rho = 1$*, the series may converge or diverge.*

APPENDIX B

MATLAB PROGRAM CALCULATING SCALING FUNCTION BY DAUBECHIES-LAGARIAS ALGORITHM

As a rule, in compactly supported orthogonal wavelets, wavelet and scaling function have no closed form. We give the MATLAB program based on Daubechies-Lagarias algorithm (Daubechies and Lagarias, ([18], [19])), that calculates a value of the scaling function at an arbitrary design point with a prescribed precision.

```
function yy = Phijk(z, j, k, filter, n)
%------------------------------------------------------------
% yy=Phijk(z, j, k, filter, n)
% Evaluation of the scaling function corresponding to an Orthogonal
% MRA by Daubechies-Lagarias Algorithm.
% inputs: z -- the argument
% j -- scale
% k -- shift
% filter -- ON finite wavelet filter, might be an
% output of WaveLab's: MakeONFilter
% n -- precision of approximation maesured by the number
% of Daubechies-Lagarias steps (default n=20)
%------------------------------------------------------------
% output: yy -- value of father wavelet (j,k) coresponding to
% 'filter' at z.
%------------------------------------------------------------------
% Example of use:
% > xx = 0:0.01:6; yy=[];
% > for i=1:length(xx)
% > yy =[yy Phijk(x(i), 0, 1, MakeONFilter('Daubechies',4), 25)];
% > end
% > plot(x,yy)
%------------------------------------------------------------------
if (nargin == 4)
n=20;
end
daun=length(filter)/2;
N=length(filter)-1;
x=(2^j)*z-k;
if(x<=0|x>=N) yy=0;
else
int=floor(x);
dec=x-int;
dy=dec2bin(dec,n);
t0=t0(filter);
t1=t1(filter);
prod=eye(N);
for i=1:n
if dy(i)==1 prod=prod*t1;
else prod=prod*t0;
end
end
```

```
y=2^(j/2)*prod;
yyy = mean(y');
yy = yyy(int+1);
end
%--------------------functions needed---------------------------
%---------
function a = dec2bin(x,n)
a=[];
for i = 1:n
if(x <= 0.5) a=[a 0]; x=2*x;
else a=[a 1]; x=2*x-1;
end
end
%-----------
function t0 = t0(filter)
%
n = length(filter); nn = n - 1;
%
t0 = zeros(nn); for i = 1:nn
for j= 1:nn
if (2*i - j > 0 & 2*i - j <= n)
t0(i,j) = sqrt(2) * filter( 2*i - j );
end
end
end
%------------------
function t1 = t1(filter)
%
n = length(filter); nn = n - 1;
%
t1 = zeros(nn); for i = 1:nn
for j= 1:nn
if (2*i -j+1 > 0 & 2*i - j+1 <= n)
t1(i,j) = sqrt(2) * filter( 2*i - j+1 );
end
end
end
%---------------------- B. Vidakovic, 2002 --------------------
%--------------------------------------------------------------
function yy = Psijk(z, j, k, filt, n)
%--------------------------------------------------------------
% yy=Psijk(z, j, k, filter, n)
% Evaluation of the wavelet function corresponding to an Orthogonal
% MRA by Daubechies-Lagarias Algorithm.
% inputs: z -- the argument
% j -- scale
% k -- shift
% filter -- ON finite wavelet filter, might be an
% output of WaveLab's: MakeONFilter
% n -- precision of approximation maesured by the number
% of Daubechies-Lagarias steps (default n=20)
%-------------------------------------------------------
% output: yy -- value of mother wavelet (j,k) coresponding to
% 'filter' at z.
%--------------------------------------------------------------
% Example of use:
% > xx = -1:0.01:5; yy=[];
% > for i=1:length(xx)
% > yy =[yy Psijk(x(i), 0, 1, MakeONFilter('Daubechies',4), 25)];
% > end
% > plot(x,yy)
%--------------------------------------------------------------
if (nargin == 4) n=20;
end
N=length(filt)-1;
```

```
daun = (N+1)/2;
x=(2^j)*z-k;
if(x<=daun-N|x>=daun) yy=0;
%if(x<=0|x>=N) yy=0;
else
twox = 2 * x;
inti=floor(twox);
dec=twox-inti;
dy=dec2bin(dec,n);
t0=t0(filt);
t1=t1(filt);
prod=eye(N);
for i=1:n
if dy(i)==1
prod=prod*t1;
else prod=prod*t0;
end
end
uu=[];
for i=1:N
index = i + 1 - inti;
if ( index > 0 & index < N + 2 )
fi= (-1)^(-index)*filt(index);
else fi=0;
end
uu =[uu fi];
end
%----------------------------------------
v = 1/N * ones(1,N) * prod' ;
yy=2^(j/2)* uu * v';
end
%--------------------functions needed----------------------------
%---------
function a = dec2bin(t,n)
a=[];
for i = 1:n
if(t <= 0.5) a=[a 0]; t=2*t;
else a=[a 1]; t=2*t-1;
end
end
%-----------
function t0 = t0(filt)
%
n = length(filt); nn = n - 1;
%
t0 = zeros(nn); for i = 1:nn
for j= 1:nn
if (2*i - j > 0 & 2*i - j <= n)
t0(i,j) = sqrt(2) * filt( 2*i - j );
end
end
end
%------------------
function t1 = t1(filt)
%
n = length(filt); nn = n - 1;
%
t1 = zeros(nn); for i = 1:nn
for j= 1:nn
if (2*i -j+1 > 0 & 2*i - j+1 <= n)
t1(i,j) = sqrt(2) * filt( 2*i - j+1 );
end
end
end
%--------------------- B. Vidakovic, 2002 --------------------
```

APPENDIX C

NANO-SCALE IMAGE ANALYSIS (NSIA) MATLAB TOOLBOX MANUAL

The method is implemented in the MATLAB toolbox NSIA. This manual provides installation and step-by-step instructions for the toolbox.

C.1 Installation

In order to install MATLAB toolbox `NSIA`, unzip all the required files into your MATALAB "work" folder (for example `c:\matlab6p5\work\nsia`). Next, create a temporary folder `c:\matlab6p5 \work\nsia\temp`. The name of the MATLAB root folder can be arbitrary, however the structure `\work\nsia` and `\work\nsia\temp` must be preserved. Finally, add a path to MATLAB. This can be done by opening MATLAB and going to `File`. Click onto `Set Path...`, and then `Add Folder...`. Select your NSIA folder, and save. This last step can be avoided if one makes `c:\matlab6p5\work\nsia` the "current directory" every time before starting the program.

C.2 Getting Started

C.2.1 Main Menu (Figure 33).

Figure 33 shows the start menu of the program. It consists of six buttons:

1. "`Select Image`": This button opens the "Image Selection" window (shown in Figure 34) which allows one to open an image file and select a region of interest.
2. "`Hough Transform`": This button opens the "Hough Transform" window (shown in Figure 35) which performs the Hough transformation of the selected image.
3. "`Analysis`": This button opens the "Analysis" window (shown in Figure 36) which detects various orientations and estimates the distance between lines formed by an atomic lattice.
4. "`Convert to meters`": This button opens the "Convert to meters" window (shown in Figure 37) which allows the conversion of average distance detected in the analysis stage to meters.
5. "`CDWT`": This button opens the "CDWT" window (shown in Figure 38) which performs a continuous directional wavelet transformation along the detected direction/orientation of the selected image.
6. "`Close`": This button closes the application.

C.2.2 "Image Selection" Window (Figure 34).

1. In this field the image is displayed.
2. This filed displays the selected region.
3. This allows the selection the size of the subimage in a pop-up menu. The "`Select`" button displays the crosshair which allows for the selection of a region of interest.
4. "`Open Image`" starts the standard menu for opening files. It allows the selecting of JPG image from files on the hard drive. The "`Save`" button saves the selected subimage. The "`Close`" button closes the current window.

C.2.3 "Hough Transform" Window (Figure 35).

1. The "`Load`" button loads the last saved subimage.
2. This field displays an image.
3. The "`Half Degree Step`" allows one to select $\Delta\theta = 0.5$ degrees in the Hough transform. This will increase sensitivity as well as computational time. The "`Hough Transform`" button performs the Hough transformation of the image.
4. This field displays the Hough transformation of the image.
5. The "`Save`" button saves the information required for the next step. The "`Close`" button closes the current window.

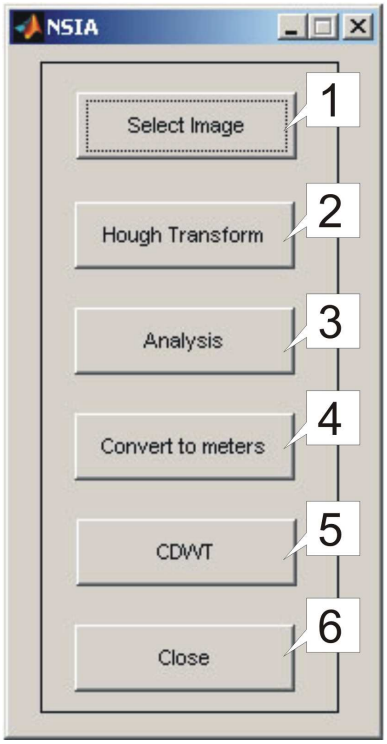


Figure 33: Main menu.

C.2.4 "Analysis" Window (Figure 36).

1. The "`Load`" button loads the last selected image. The "`Close`" button closes the current window and saves the results required for the next step.
2. The "`Plot Image`" button displays the image in field (7).
3. The controlled parameters field allows to change parameters to increase sensitivity of the method.
4. The "`Detect`" button performs the analysis of the image and detects parallel lines formed by an atomic lattice, and measures the distance between them. The results of the analysis are displayed in (5) and (6).
5. This displays the plot of the energy function, with circles representing detected angles/orientations of the parallel lines.
6. This field lists the results of the analysis (detected orientation, number of parallel lines, average distance between the lines and standard deviation of the distances). Selecting particular orientation and clicking the "`Plot`" button will plot lines of this orientation over the image in field (7), graph of the distances between the lines in field (8), and display the histogram of the distances in the field (9).
7. This field displays the image under investigation.
8. This field displays the graph of the distances between the lines.
9. This field displays the histogram of the distances between the lines.
10. This is the control over calculation of the average distance between the lines. There are three possible choices in the pop-up menu: standard average, trimmed average, and winsorized average. Upper and lower quantiles for the trimmed and winsorized averages are shown in corresponding fields. Select the desired method in the menu, enter the values of the quantiles if necessary, and click the "`Recalculate Avg.`" button to get the results.
11. This is the control over the elimination of the close lines. Select the "`Eliminate Close Lines`" option together with the number of pixels in the pop-up menu, and click the "`Detect`" button to get the results.

C.2.5 "Convert to meters" Window (Figure 37).

1. The "`Load`" button loads the results from the previous step of the analysis. The "`Calculate`" button converts the results to meters.

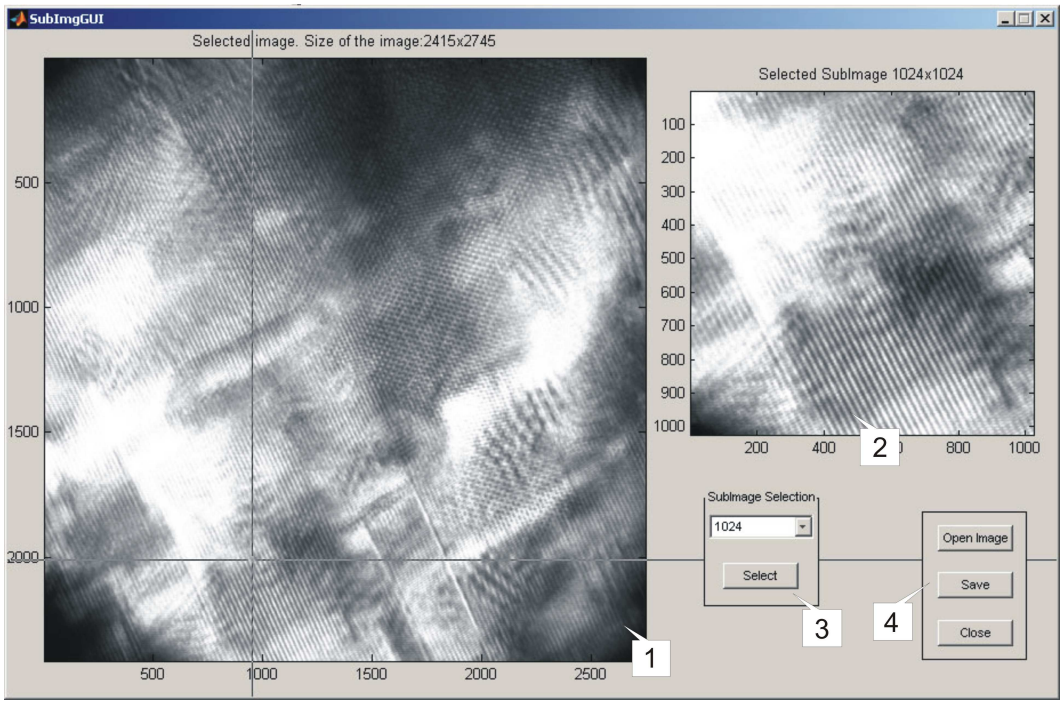


Figure 34: `Select Image` window.

2. This menu allows the selection of the resolution at which the image was originally scanned.
3. This menu allows the selection the magnification of the microscope used.
4. This field displays the results.

C.2.6 "CDWT" Window (Figure 38).

1. The "`Load`" button loads the image under investigation (displayed in field (4)). The "`Save`" button saves the results of the continuous directional wavelet transformation on the hard drive. The "`Close`" button closes current window.
2. This is a list of all detected orientations. Select the desired orientation by clicking on it.
3. Select the scale of the wavelet transformation by sliding the rule or entering the value in a field. The "`CDirWT`" button performs the transformation.
4. This field displays the image under investigation.
5. This field displays the results of the wavelet transform.

C.3 Step-by-step Instructions.

- Type "nsia" in the MATLAB command line (`>>nsia`) in order to start the program. Figure 33 shows the start menu. To make a full analysis one needs to go through each of the steps.
- Analysis starts with the loading of the image and selection of a subimage of interest. Click `Select Image` button (Figure 33.(1)). "Image Selection" window shown in Figure 34 will appear.
- Click the `Open Image` button (Figure 34.(4)). The standard menu for opening file will appear (currently the program works only with JPG format). Select the desired file. The field Figure 34.(1) will display the selected image.
- Select the subimage size in the pop-up menu Figure 34.(3). The available sizes are 256×256, 512×512, 1024×1024, 2048×2048 and 4096×4096. The larger the subimage size, the longer it will take to run the analysis. The recommended subimage size is 1024×1024.

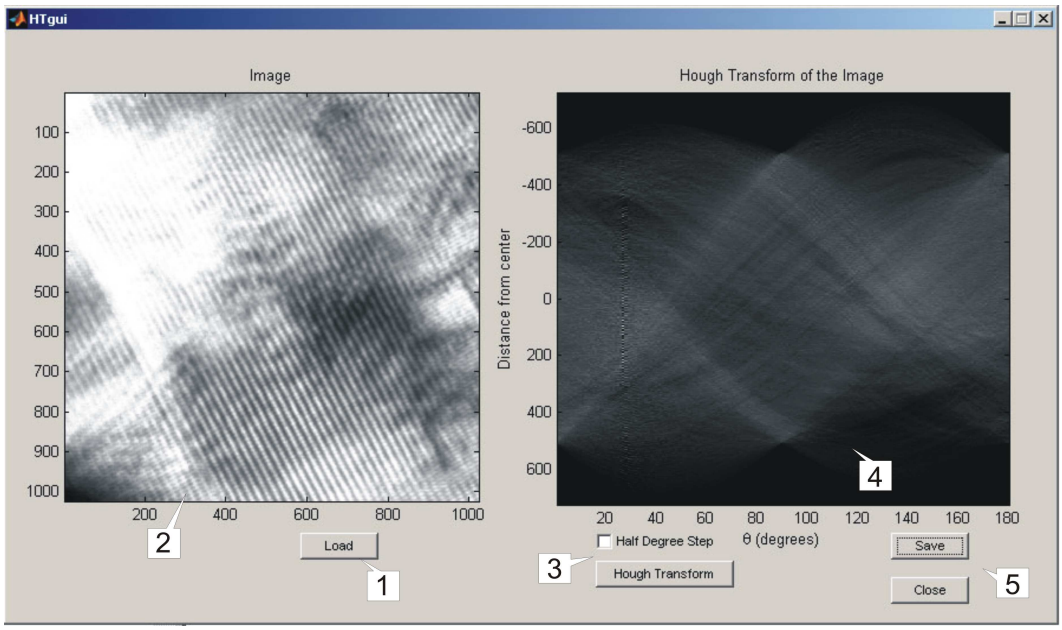


Figure 35: `Hough Transform` window.

- Click the `Select` button Figure 34.(3). The crosshair appears, click on the region of interest. The filed Figure 34.(2), will display the selected region. This step can be repeated until desired region is selected.
- Then satisfied with the selected subimage, click `Save` and `Close` buttons (Figure 34.(4)). This will save the selected subimage and close this window.
- Click the `Hough Transform` button (Figure 33.(2)). The "Hough Transform" window shown in Figure 35 will appear.
- Clicking on the `Load` button (Figure 35.(1)) the last selected subimage will be loaded and shown in (Figure 35.(2)).
- Now one has an option of selecting "Half degree step" (Figure 35.(3)). This will increase sensitivity as well as calculation time. It is not recommend to select this option during the first run. If the procedure, however, fails to detect orientations, one can select this option in a second run. Click the `Hough Transform` button. This is the longest computational step in the program. Upon completion, the Hough transform of the image will be displayed in (Figure 35.(4)).
- Click the `Save` and `Close` buttons to save the results and close the window (Figure 35.(5)).
- Click the `Analysis` button (Figure 33.(3)). The "Analysis" window shown in Figure 36 will appear.
- Click the `Load` button (Figure 36.(1)). This loads the previously saved information.
- Click the `Plot Image` button (Figure 36.(2)) to display the image under investigation.
- Click the `Detect` button (Figure 36.(4)). For the first run it is recommended to run the analysis with the default parameters. A graph will appear in Figure 36.(5) with circles representing detected angles . Information about the detected angles, number of detected lines in each direction as well as the average distance between the lines of the same orientation will be listed in the "Results list" Figure 36.(6).
- One can select any detected direction and using the `Plot` button (Figure 36.(6)), draw all the lines of the selected direction. A graph which shows the distances (Figure 36.(8)) as well as the histogram of the distances (Figure 36.(9)) will also appear.
- One can change the controlled parameters in order to detect additional directions by sliding the bars or entering the parameter value manually (Figure 36.(3)). After changing the parameters, click the `Detect` button again.
- It is also possible to change the way average distance are calculated in the pop-up menu.Usual average, trimmed average (enter the lower and upper percentages for the trimming), or winsorized average can be calculated (Figure 36.(10)).
- When all desired orientation are found, it is recommended to improve the results by elimination of the close lines. The reason for this is that in the real life images the lines are almost never straight and always discontinuous.

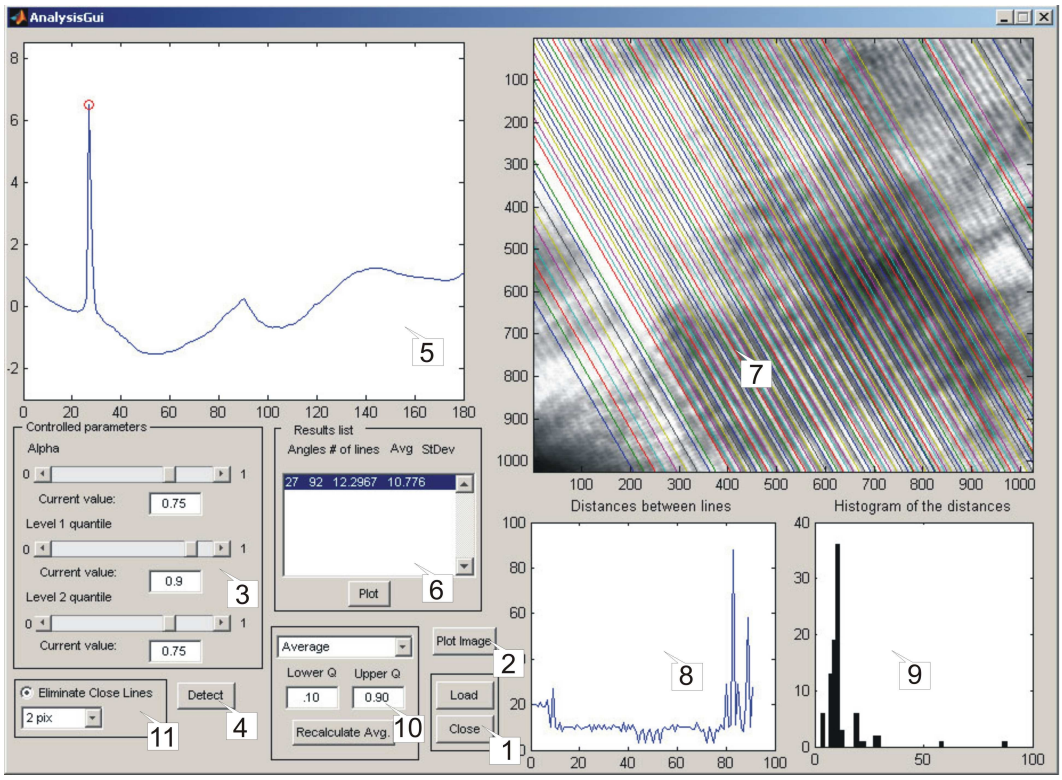


Figure 36: `Analysis` window.

This creates the problem of close lines, when a single line generates several broken segments which are getting detected as several lines. This introduces large errors into the analysis of the distances between the lines. The elimination of the closes can be done in two steps. First select `Eliminate Close Lines` (Figure 36.(10)). Next, in a menu below select 3pix, 2pix or 1pix, and click the `Detect` button again. This step takes some time. It will through all detected lines of all detected directions, finding the lines with distances between which are less that 3 pixels, 2 pixel or 1 pixel, eliminating insignificant ones. This step significantly improves results.

- Close the window by clicking the `Close` button. It will automatically save information required for the next step.
- To conversion of distances from pixels to meters. Click the `Convert to meters` button (Figure 33.(4)). The "Convert to meters" window shown in Figure 37 will appear.
- Click the `Load` button (Figure 37.(1)).
- Select the resolution at which the image was originally scanned in the pop-up menu (Figure 37.(2)), and select the magnification of the microscope used (Figure 37.(3)).
- Click `Calculate` to get the results.
- The final step is CDWT, which stands for Continuous Directional Wavelet Transformation. This feature will help to see the structure of layers. Click the `CDWT` button (Figure 33.(5)).
- Click the `Load` button (Figure 38.(1)). The image under investigation will appear in Figure 37.(4)), as well as all detected directions in Figure 37.(2).
- Select the desired direction from the list (Figure 37.(2)). Select the scale of the wavelet transformation (Figure 37.(3)). The recommend the scale to is around 4-8.
- Click on the `CDirWT` button (Figure 37.(3)) to get the results. The filed in Figure 37.(5) will display the wavelet transformation along the specified direction.
- Click on the `Save` button, to save the results. The results will be saved in CWTDdata.mat file in `...\nsia\temp` folder.

C.4 Comments

The following functions are from the WaveLab toolbox for the MATLAB.

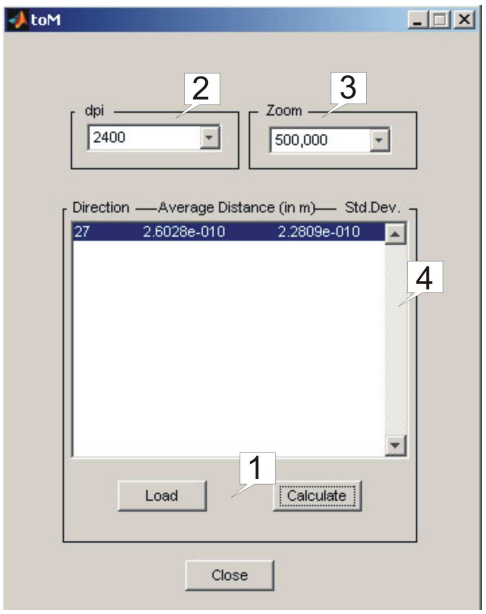


Figure 37: `Convert to meters` window.

Available at `http://www-stat.stanford.edu/~wavelab/`.

ancov.m, downdyadhi.m, downdyadlo.m, dyad.m, dyadlength.m, fwt_po.m, fwt_stat.m, fwt_ti.m, iconv.m iwt_po.m, iwt_stat.m lshift.m, makeonfilter.m, mirrorfilt.m, packet.m, reverse.m, rshift.m, shapeasrow.m, shapelike.m, ti2stat.m, updyadhi.m, updyadlo.m, upsample.m

And the following functions are from YAWTB toolbox.

Available at `http://www.fyma.ucl.ac.be/projects/yawtb/index.php`.

cauchy2d.m, cwt_2d.m, getopts.m, list_elem.m, yapuls.m, yawopts.m.

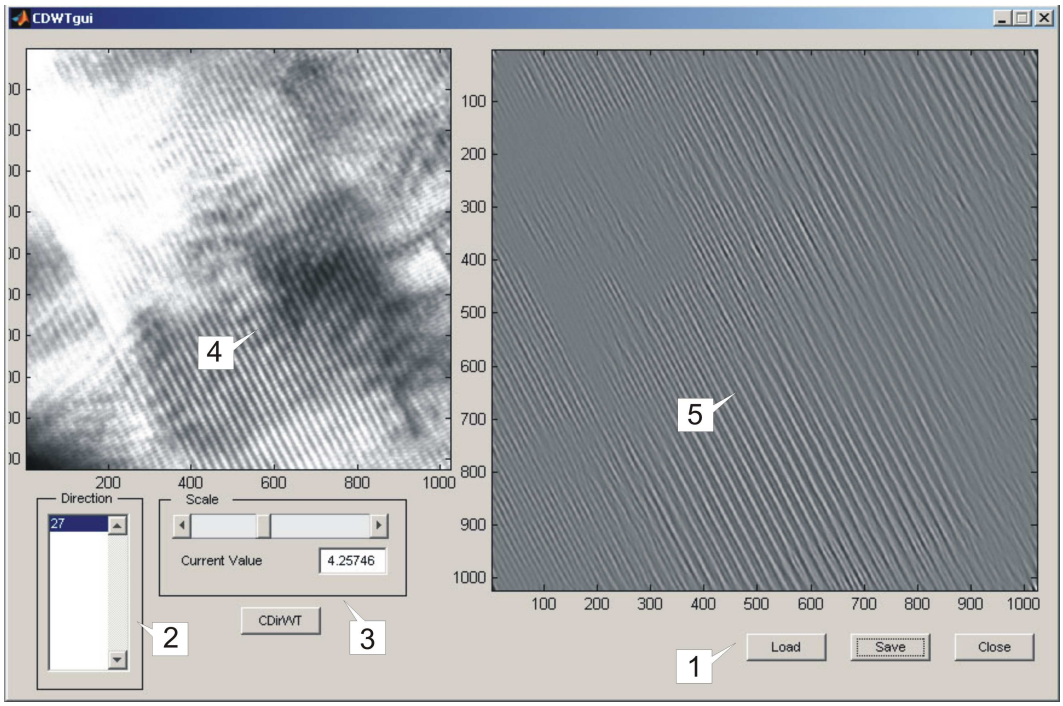


Figure 38: `CDWT` window.

REFERENCES

[1] ABRAMOVICH, F., BAILEY T., and SAPATINAS T. (2000) *Wavelet analysis and its statistical allications*, The Statistician, **49** Part 1, 1–29.

[2] ABRAMOVICH, F. and BENJAMINI, Y. (1995) *Thresholding of wavelet coefficients as a multiple hypotheses testing procedure*, Wavelets and Statistics, Editors A. ANTONIADIS and G. OPPENHEIM. Lecture Notes in Statistics, **103**, 5–14. Springer-Verlag, New York.

[3] ABRAMOVICH, F. and BENJAMINI, Y. (1996) *Adaptive thresholding of wavelet coefficients*, Computational Statistics and Data Analysis, **22** 351–361.

[4] ABRAMOVICH, F. and SILVERMAN, B.W. (1998) *The vagulette-wavelet decomposition approaches to statistical inverse problem*, Biometrika, **85**, 115–129.

[5] ANGELINI, C. and VIDAKOVIC, B. (2003) *Some Novel Methods in Wavelet Data Analysis: Wavelet Anova, F-test Shrinkage, and* Γ*-Minimax Wavelet Shrinkage*, Wavelets and their Applications, Editors KRISHNA, M., RADHA, R., and THANGAVELY, S. Allied Publishers Ltd, New Delhi, ISBN 81-7764-493-9, pp. 31–45.

[6] ANTONIADIS, A., BIGOT, J. and SAPATINAS, T. (2001) *Wavelet estimators in nonparametric regression: a comparative simulation study*, Journal of Statistical Software, **6**, 1–83.

[7] ANTONIADIS, A. and GIJBELS, I. (2000) *Detecting abrupt changes by wavelet methods*, J. Nonparam. Statist., **10**.

[8] ANTONIADIS, A. and OPPENHEIM, G. (editors) (1995) *Wavelets and Statistics*, Lecture Notes in Statistics, **103**, Springer-Verlang.

[9] BAILEY, T.C., SAPATINAS, T., POWELL, K.J. and KRZANOWSKI, W.J. (1998) *Signal detection in underwater sound using wavelet transforms*, Technical Report 98-01. Computational Mathematics Laboratory, Rice University, Houston.

[10] BRIGHAM, E. O. (1988) *The Fast Fourier Transform and Its Applications,* Prentice-Hall, Englewood Cliffs, NJ.

[11] BENJAMINI, Y. and HOCHBERG, Y. (1995) *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, J.R. Stat. Soc. Ser. B Stat. Methodol. **57**, 289–300.

[12] BINNIG, G., QUATE, C.F., and GERBER, CH. (1986). *Atomic force microscope*. Phys. Rev. Lett. **56**, 930–933.

[13] CHANG, W. KIM, S. and VIDAKOVIC, B. (2003) *Wavelet-Based Estimation of a Discriminant Function.* Applied Stochastic Models in Business and Industry, 19 185–198.

[14] CHIPMAN, H., KOLACZYK, E., and MCCULLOCH, R. (1997) *Adaptive Bayesian Wavelet Shrinkage*, Journal of the American Statistical Association, **92**, 1413–1421.

[15] COIFMAN, R.R. and DONOHO, D.L. (1994) *Translation Invatriant denoising* in *Wavelets and Statistics* (ed. A. Antoniadis and G. Oppenheim). Springer Lecures Notes in Statistics **103**, 125–150.

[16] DATTA, S. and DATTA, S. (2005) *Empirical Bayes screening of many p-values with applications to microarray studies*. Bioinformatics, **21**, 1987–1994.

[17] DAUBECHIES, I. (1992) *Ten Lectures on Wavelets*, Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM.

[18] DAUBECHIES, I. and LAGARIAS, J. (1991) *Two-scale difference equations I. Existence and global regularity of solutions*, SIAM J. Math. Anal., **22**, 5, 1388–1410.

[19] DAUBECHIES, I. and LAGARIAS, J. (1992) *Two-scale difference equations II. Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., **23**, 4, 1031–1079.

[20] DEANS, S.R. (1981) *Hough Transform From the Radon Transform*, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-3(2).

[21] DEVROYE, L. GYÖRFI, L. and LUGOSI, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, NY.

[22] DONOHO, D. L. and JOHNSTONE, I. M. (1994) *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, **81**, 425–455.

[23] DONOHO, D. L. (1995) *De-noising by soft thresholding*, IEEE Transactions on Information Theory, **41**, 613–627.

[24] DUDA R.O., and HART P.E., (1972) *Use of the Hough Transform to Detect Lines and Curves in Pictures*, Comm ACM 15, pp 11-15.

[25] DUDOIT, S., SHAFFER, J., and BOLDRICK, J. (2003) *Multiple hypothesis testing in microarray experiments*, Statistical Science,**18**, 71–103.

[26] EFRON, B. (2004) *Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis* Journal of the American Statistical Association, **99**, 96–104.

[27] EFRON, B. and TIBSHIRANI, R. (2002) *Empirical Bayes methods and false discovery rates for microarrays*, Genetic Epidemiology, **23**, 70–86.

[28] EFRON, B., TIBSHIRANI, R., STOREY, J. and TUSHER, V. (2001) *Empirical Bayes analysis of a microarray experiment*, J. Amer. Statist. Assoc. **96**, 1151–1160.

[29] FOLKS, J. L. (1984) *Combination of Independent Tests*, Handbook of Statistics, Vol 4., Editors P. R. KRISHNIAH and P. K. SEN, Elsevier Scince Publishers, 113–121.

[30] GAO, H,-Y. (1997) *Choise of threshold for wavelet shrinkage estimation of the spectrum* J. Time Ser. Anal., **18**, 231–251.

[31] GREBLICKI, W. (1981) *Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities*, IEEE Transactions on Information Theory, **27,** 3, 364–366.

[32] GREBLICKI, W. and RUTKOWSKI, L. (1981) *Density-free Bayes risk consistency of nonparametric patters recognition procedures*, Proceedings of the IEEE, **69,** 4, 482–483.

[33] GREBLICKI, W. and PAWLAK, M. (1982) *A classification procedure using the multiple Fourier series*, Information Sciences, **26,** 115–126.

[34] GREBLICKI, W. and PAWLAK, M. (1983) *Almost sure convergence of classification procedures using Hermite series density estimates*, Pattern Recognition Letters, **2,** 13–17.

[35] GROSSMANN, A. and MORLET, J., (1984) *Decomposition of Hardy functions into square integrable wavelets of constant shape,* SIAM J. Math., 15:723–736.

[36] GROSSMANN, A. and MORLET, J., (1985) *Decomposition of functions into wavelets of constant shape and related transforms,* In L. Streit, editor, *Mathematics and physics, lectures on recent results*. World Scientific, River Edge, NJ.

[37] HOUGH P.V.C. (1962) *Method and Means for Recognizing Complex Patterns*, U.S. Patent No. 3069654.

[38] ILLIGWORTH J. AND KITTLER J. (1988) *A Survey of the Hough Transform*, Computer Vision Graphics and Image Processing, 44, pp 87-116.

[39] JOHNSTONE, I. M. and DONOHO, D. L. (1995) *Adapting to smoothness via wavelet shrinkage*, Journal of the Statistical Association, **90**, 1200–1224.

[40] JUNG, Y.-Y., OH, M.-S., SHIN, D. W., KANG, S.-H., and OH, H. S. (2005) *Identifying direrentially expressed genes in meta-analysis via Bayesian model-based clustering*

[41] KOHLER, M. (2001) *Nonlinear orthogonal series estimates for random design regression*. Technical Report, Department of Mathematics, University of Stutgart, Germany. `http://www.mathematik.uni-stuttgart.de/mathA/lst3/kohler/papers-en.html`

[42] LEAVERS V.F. (1992) *Shape detection is Computer Vision Using the Hough Transform*, Springer-Verlag.

[43] JUN LU, GAD GETZ, ERIC A. MISKA, EZEQUIEL ALVAREZ-SAAVEDRA, JUSTIN LAMB, DAVID PECK, ALEJANDRO SWEET-CORDERO, BENJAMIN L. EBERT, RAYMOND H. MAK, ADOLFO A. FERRANDO, JAMES R. DOWNING, TYLER JACKS, H. ROBERT HORVITZ and TODD R. GOLUB (2005) *MicroRNA expression profiles classify human cancers*, Nature, 435(9), 834–838.

[44] MAGLI E., and OLMO G. (2000) *Integrated Compression and Linear Feature Detection in the Wavelet Domain*, ICIP 2000 – IEEE International Conference on Image Processing, Vancuver, Canada.

[45] MALLAT, S. (1989) *Multiresolution approximations and wavelet orthonormal bases of* $\mathbb{L}^2(\mathbb{R})$, Trans. Amer. Math. Soc., 315:69–87.

[46] MALLAT, S. (1989) *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Trans. on Patt. Anal. Mach. Intell., 11(7):674–693,.

[47] MALLAT, S. (1997) *A Wavelet Tour of Signal Processing*, Academic Press.

[48] MALLAT, S. and HWANG W. (1992) *Singularity detection and precessing with wavelets*, IEEE Trans. Inf. Theory, **38**, 617–643.

[49] MARSHALL, B., MCEVER, R. and ZHU, C. (2001) *Kinetic rates and their force dependence on the P-Selectin/PSGL-1 interaction measured by atomic force microscopy*, Proceedings of ASME 2001, Bioengineering Conference, BED - Vol. 50.

[50] MILLER, R. (1981) *Simultaneous Statistical Inference*, Second Edition. Springer-Verlag, NY.

[51] MORLET, J., ARENS, G., FOURGEAU, E., GIARD, D. (1982) *Wave propagation and sampling theory*, Geophys. **47**, 203–236.

[52] MORTTIN, P.A. (1996) *From Fourier to wavelet analysis of time series*, In Proc. Computational Statistics (ed. A. Pratt), 111–122, New York: Physica.

[53] MOULIN, P. (1993) *Wavelet thresholding thechniques for power spectrum estimation*, IEEE Trans. Signal Process., **42**, 3126–3136.

[54] NASON, G.P. and VON SACHS, R. (1999) *Wavelets in time series analysis* Phil. Trans. R. Soc. Lond. A, **357**, 2511–2526.

[55] NASON, G.P. and SILVERMAN, B.W. (1995) *The stationary wavelet transform and some statistical applications.* in *Wavelets and Statistics* (ed. A. Antonidis and G. Oppenheim). Springer Lecture Notes in Statistics **103** 281–300.

[56] NASON, G.P., VON SACHS, R. and KROISANDT, G. (2000) *Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum*, J. R. Statist. Soc. B, **62**, 271-292.

[57] NEUMANN, M.H. (1996) *Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series*, J. Time Ser. Anal., **17**, 601–633.

[58] NEUMANN, M.H. and VON SACHS, R. (1997) *Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra*, Ann. Statist., **25**, 38-76.

[59] OGDEN, T. (1997) *Essential Wavelets for Statistical Applications and Data Analysis*, Boston:Birkhäuser.

[60] OGDEN, T. and LYNCH, J. (1999) *Bayesian analysis of change-point models*, Lect. Notes Statist., **141**, 67–82.

[61] OGDEN, T. and PARZEN, E. (1996) *Data dependent wavelet thresholding in nonparametric regression with change-point applications*, Computational Statistics and Data Analysis, **22**, 53–70.

[62] PANDIT, S. and WU, S-M. (1993) *Time Series and System Analysis with Applications.* Krieger Publishing Company, Malabar, Fl.

[63] PERCIVAL, D.B. and WALDEN, A.T. (1999) *Wavelets Methods for Time Series Analysis*, Cambridge: Cambridge University Press.

[64] PRIESTLEY, M.B. (1999) *Wavelets and time-dependent spectral analysis*, J. Time Ser. Anal., **17**, 85–104.

[65] RAIMONDO, M. (1998) *Minimax estimation of sharp change points*, Ann. Statist., **26**, 1379–1397.

[66] RICHWINE, J. (1996) *Bayesian estimation of change-points using Haar wavelets* MSc Thesis. Department of Statisitcs, University of South Carolina, Columbia.

[67] ROSNER, G. and VIDAKOVIC, B. (2000) *Wavelet Functional ANOVA, Bayesian False Discovery Rate, and Longitudinal Measurements of Oxygen Pressure in Rats*, Technical Report 1/2000, ISyE, Georgia Institute of Technology.

[68] VON SACHS, R. and SCHNEIDER, K. (1996) *Wavelet smoothing of evolutionary spectra by nonlinear thresholding*, Appl. COmput. Harm. Anal., **3**, 268–282.

[69] TADESSE, M., IBRAHIM, J., VANNUCCI, M., and GENTLEMAN, R. (2005) *Wavelet Thresholding with Bayesian False Discovery Rate Control*, Biometrics,**61**, 25–35.

[70] VAN RYZIN, J. (1966) *Bayes risk consistency of classification procedures using density estimates*, Sankhyā, Ser. A 28, 161–170.

[71] VANDERGHEYNST P. and GOBBERS J-F. (2002) *Directional Dyadic Wavlet Transforms: Design and Algorithms* IEEE Transactions on Image Processing, **11**, 4, 363–372.

[72] VIDAKOVIC, B. (1998) *Nonlinear wavelet shrinkage with Bayes rules and Bayes factors* Journal of the American Statistical Association, **93**, 173-179

[73] VIDAKOVIC, B. (1999) *Statistical Modeling by Wavelets.* John Wiley & Sons, Inc., New York, 384 pp.

[74] VIDAKOVIC, B. AND RUGGERI, F. (2001) BAMS Method: Theory and Simulations. *Sankhyā, Series B,* **63***,2* (Special Issue on Wavelets), 234–249.

[75] WANG, Y. (1995) *Jump and sharp cusp detection by wavelets*, Biometrika, **82**, 385–397.

[76] WOJTASZCZYK, P. (1997) *A Mathematical Introduction to Wavelets*, London Mathematical Society Student Texts, **37**.

INDEX

VITA

Ilya Lavrik received the B.S. degree in Mathematics from the Syktyvkar State University, Russia, in 1999. He received M.S. degrees in Applied Mathematics and Statistics in 2001 and 2002. From 2002 he is a Ph.D. student in the School of Industrial and Systems Engineering at Georgia Institute of Technology.